Chapter 3. Results

This chapter presents the results of our review, organized around the key questions.

Question 1: What are the test characteristics (reliability, sensitivity, specificity, predictive values) and costs of measures used in the management of prolonged pregnancy to (a) assess risks to the fetus and mother of prolonged pregnancy, and (b) assess the likelihood of a successful induction of labor?

Approach

Assessment of Risks to Fetus and Mother

In Chapter 1, we discussed the evidence for increasing risk of adverse outcomes, especially perinatal death, as gestational age advances beyond 40 weeks. Although this risk is small in absolute terms, the trend towards increasing risk with increasing gestational age is consistent across studies. One approach to preventing these adverse outcomes would be to use testing to identify patients most likely to experience them.

Which antenatal testing strategies lead to improvements in fetal and maternal outcomes? The best way to answer this question is with studies that directly compare one testing strategy with another (or no testing), with the least biased assessment from a randomized control trial, followed by concurrent nonrandomized cohort comparisons, historical cohort comparisons, and cohort studies with variation in testing strategies employed (Evidence Table 1).

However, most of the published literature consists of case series or cohort studies in which there is little or no variation in testing strategies (or variation is not reported). Such studies are less useful but still may contain valuable information concerning the association of test results with fetal and maternal outcomes.

This association can take one of two forms, either prediction of future outcomes (for example, association of antenatal nonstress test [NST] with low Apgar scores or neonatal mortality) or assessment of current status (e.g., measuring abdominal circumference in utero by ultrasound to assess incidence of macrosomia or fetal weight). These studies address the question, “How accurate is the assessment of current fetal status or prediction of future maternal and fetal outcomes offered by antenatal testing?” While evidence that one test is more accurate or has a stronger association with relevant outcomes suggests that it would be more effective, this is by no means definitive. Nevertheless, most of the studies providing data about the predictive value of the tests considered provided 2-by-2 table data (Table 6).

Reliability of Tests

We additionally sought data on the reliability of tests, including interobserver variation, when these were available. If a test result is not reproducible when the test is performed by different examiners, or by the same examiner on different occasions, then the utility of the test is reduced, even if the “average” test characteristics (sensitivity, specificity) imply useful discrimination or prediction.

Correlation of Tests

In certain cases, the association of one test result with another was reported without reference to outcomes.

Results

Assessment of Risks to the Fetus Associated with Uteroplacental Insufficiency

Testing versus no testing. We did not identify any randomized trials in which women with prolonged gestation were randomly assigned to antepartum surveillance or no testing. Of four randomized trials of antepartum cardiotocography versus no surveillance in "high-risk" pregnancies (Brown, Sawers, Parsons, et al., 1982; Flynn, Kelly, Mansfield, et al., 1982; Kidd, Patel, and Smith, 1985; Lumley, Lester, Anderson, et al., 1983)—also the subject of a systematic review by Pattison and McCowan (2001)—only one (Flynn, Kelly, Mansfield, et al., 1982) included patients who were being followed explicitly for prolonged gestation (classified as "suspect postmaturity syndrome" in the paper). In this trial, 100 of 300 subjects were being followed for this indication. All patients received either outpatient ("at intervals of not more than 1 week") or inpatient ("at least twice per week") NSTs. Patients were randomized to two groups: in one, clinicians taking care of the patients knew the results of the NST, while in the other group, NST results were not revealed. Although quantitative data were not reported on this, it appears that the majority of the patients with prolonged gestation received outpatient testing between 41 and 42 weeks, when induction was scheduled.

Although results were not reported separately for women with prolonged gestation, there were no statistically significant differences in stillbirths, neonatal deaths, or other adverse neonatal outcomes between the two groups. However, patients in the group in which caregivers knew the results were significantly more likely to be discharged from the hospital before delivery and significantly more likely to receive outpatient care. There also were nonsignificant trends towards fewer antenatal inpatient days and fewer elective cesarean sections in the group whose caregivers were aware of their results.

In this study (Flynn, Kelly, Mansfield, et al., 1982), a nonreactive NST had 100 percent sensitivity for stillbirths with nonlethal congenital abnormalities and a specificity of 88 percent; positive predictive value was nine percent, and negative predictive value 100 percent. None of the deaths were in the prolonged pregnancy group. Test characteristics for surrogates of fetal compromise were less favorable. For fetal distress in labor, sensitivity was 37 percent, specificity 88 percent, positive predictive value 18 percent, negative predictive value 93 percent. Similar trends were seen for meconium and admission to the neonatal intensive care unit: considerably lower sensitivity than specificity, poor positive predictive value, and good negative predictive value. These findings suggests that the effects on management observed in this trial—consistent trend towards less aggressive observational strategies in the group where the results were revealed to clinicians—reflect clinically appropriate interpretation of the test results. The high negative predictive values are evidence that a normal test does provide reassurance. Unfortunately, the paper does not allow estimation of test characteristics in the specific population of interest for this report, patients with prolonged pregnancy and no other risk factors.

We did identify two retrospective concurrent cohort studies comparing testing and no testing in women with prolonged pregnancy (Bochner, Williams, Castro, et al., 1988; Fleischer, Schulman, Farmakides, et al., 1985). Fleischer, et al., reported a retrospective cohort study comparing 228 women who had weekly NST monitoring beginning at 41 weeks with 30 women who had no antenatal monitoring (Fleischer, Schulman, Farmakides, et al., 1985). Reasons for women not receiving testing were not specified. Despite the small sample size of the no-testing group, the investigators observed significant differences in most of the outcome variables they reported, including low Apgar score (< 7) at 1 and 5 minutes, neonatal intensive care unit (NICU) admission rates, stillbirth rates, and cesarean section for fetal distress. The small sample of women with no monitoring, the retrospective nature of the study design, and the unusually high rates of adverse fetal and maternal outcomes all suggest that the no-testing group in this study may be dissimilar to the NST monitoring group in other ways besides whether an antenatal NST was conducted. This potential confounding probably exaggerates the effectiveness of NST monitoring.

Bochner, et al., described a comparison of large concurrent cohorts of women who underwent antenatal testing with amniotic fluid volume (AFV) and nonstress testing beginning at week 41 or 42 and those with no antenatal testing (Bochner, Williams, Castro, et al., 1988). They found an association with total number of adverse outcomes (testing, 0/512; no testing, 13/1807 [0.7 percent]; $p < 0.05$) and a trend toward higher cesarean section for fetal distress in the no-testing cohort (testing, 14/512 [2.7 percent]; no testing, 60/1807 [3.3 percent]; $p = 0.07$). When the results of testing were compared in the groups beginning testing at 41 weeks ($n = 908$) and those at 42 weeks ($n = 352$), the positive predictive value for a diagnosis of intrapartum fetal distress was significantly higher at 42 weeks (21.1 percent at 42 weeks vs. 11.9 percent at 41 weeks), with a concomitantly lower negative predictive value (98.5 percent at 42 weeks vs. 99.1 percent at 41 weeks). This is consistent with an overall increased risk of adverse outcomes with increasing gestational age, assuming that the sensitivity and specificity of the test are independent of gestational age (more on this below). It is unclear why the no-testing group did not receive testing, since women with "high risk factors" were excluded, and inclusion criteria required that women be seen prior to 20 weeks. Again, the possibility of confounding cannot be ruled out.

In summary, it is difficult to draw conclusions about the effectiveness of antepartum testing compared with no testing in prolonged pregnancy. The only randomized trial comparing testing with no testing is limited by a heterogeneous population (in terms of other risk factors), relatively small numbers of patients with prolonged pregnancy alone, failure to report results separately by indication for testing, and questions about the applicability of the results to current practice (Pattison and McCowan, 2001). The two nonrandomized studies identified suggest an excess risk of adverse outcomes in unmonitored pregnancies, but the failure to characterize the groups studied makes it impossible to rule out other factors as the cause of this excess risk.

Maternal sensation of fetal movement (kick counts). We identified only one study that assessed the association of maternal sensation of fetal movement with postmaturity syndrome, defined as characteristic skin changes (desquamation, leather-like consistency, little subcutaneous fat) and a "long, lean body," with a ponderal index (weight in grams x 100/length in cubic centimeters) of 2.27 or less (10^{th} percentile or less). Rayburn, et al., tested a group of 147 women at 42 weeks or more gestational age using the NST plus fetal movement charting plus urine estrogen-to-creatinine ratio (Rayburn, Motley, Stempel, et al., 1982). These tests were performed semi-weekly or weekly. If the NST was reactive (two adequate accelerations of baseline fetal heart rate [FHR] during a 20- to 40-minute period), then it was repeated on the next visit. If the NST was nonreactive, then the test was either repeated or a contraction stress test (CST) was given on the same day. Of the 147 cases studied, 32, or 22 percent, had postmaturity syndrome. However, none of the mothers recording kick counts noted reduced fetal movement (sensitivity, 0/32; specificity, 115/115 [100 percent]). The kick count measure was not useful for predicting postmaturity syndrome, with an undefined positive predictive value and negative predictive value of 78 percent. No studies documenting the reliability of this method (such as correlation between maternal sensation of movement and observed movements on ultrasound) were identified.

In summary, there are no data to suggest that maternal sensation of fetal movement is useful in predicting which infants are affected by postmaturity syndrome. There are no data at all to allow evaluation of maternal sensation of fetal movement as a predictor of other adverse outcomes associated with prolonged gestation.

Nonstress test (NST). We identified one randomized trial enrolling 287 patients comparing the NST alone with a simple biophysical profile (NST plus AFV, supplemented by estimates of fetal weight and placental function) (Arias, 1987). In this trial, 44 of 217 patients had abnormal results on antenatal testing, 14/112 in the NST alone group and 30/105 in the NST + AFV group. There were no significant differences in any outcome, including fetal distress or cesarean section for fetal distress, though slightly more inductions and cesarean sections for fetal distress occurred in the biophysical profile arm. Test characteristics of other components of this combination of tests (ultrasound for fetal weight alone, ultrasound for placental function alone, or ultrasound for AFV alone) were not reported. Sensitivity was similar for NST alone and NST + AFV; however, specificity was higher for NST alone than for NST + AFV. This study was rated positively for 9 of 12 quality assessment items, failing items for sample size and statistical analysis.

Eleven articles provided 40 separate 2-by-2 tables addressing the association of NST with intermediate fetal and maternal outcomes (Arias, 1987; Devoe and Sholl, 1983; Eden, Gergely, Schifrin, et al., 1982; Farmakides, Schulman, Winter, et al., 1988; Fleischer, Schulman, Farmakides, et al., 1985; Phelan, Platt, Yeh, et al., 1984; Ramrekersingh-White, Farkas, Chard, et al., 1993; Small, Phelan, Smith, et al., 1987; Tongsong and Srisomboon, 1993; Weiner, Farmakides, Schulman, et al., 1994; Weiner, Reichler, Zlozover, et al., 1993). The outcomes considered were intermediate in six cases, fetal in 29, and maternal in five cases. The number of specific outcomes is shown in Table 7.

Table 8 shows the sensitivity and specificity, as well as positive and negative predictive values, for each study. For predicting 1-minute Apgar scores < 7, data from five studies (Eden, Gergely, Schifrin, et al., 1982; Fleischer, Schulman, Farmakides, et al., 1985; Phelan, Platt, Yeh, et al., 1984; Small, Phelan, Smith, et al., 1987; Tongsong and Srisomboon, 1993) showed that the sensitivity of NST ranged from 0.12 to 0.41, and specificity ranged from 0.81 to 0.97. For predicting low 5-minute Apgar scores, data from the same five studies and one more (Devoe and Sholl, 1983) showed that the sensitivity of NST ranged from 0 to 0.5, and specificity ranged from 0.80 to 0.95. Two studies used combined endpoints and found that NST was predictive, with sensitivity of 0.08 to 0.33 and specificity of 0.91 to 0.95.

In addition to data on the NST as a whole, two studies reported the predictive value of fetal heart rate monitoring in the context of nonstress testing (Rayburn, Motley, Stempel, et al., 1982; Sherer, Onyeije, Binder, et al., 1998) (Table 9). Neither bradycardia nor tachycardia alone had high sensitivity or specificity for predicting low Apgar scores, meconium aspiration, or NICU admission. Neither was abnormal heart rate associated significantly with the occurrence of postmaturity syndrome.

In summary, results of these studies suggest that a reactive nonstress test in prolonged pregnancy has good negative predictive value—i.e., adverse outcomes are unlikely to occur in the setting of a reactive nonstress test—but that the positive predictive values are low. Data from the one randomized trial comparing weekly NST beginning beyond 40 weeks to NST and amniotic fluid assessment suggest equivalent outcomes.

Contraction stress test (CST) using oxytocin. Knox, et al., compared the CST using oxytocin with amniocentesis for meconium staining in 187 women at 42 weeks gestation (Knox, Huddleston, and Flowers, 1979). The study was prospective, with women assigned to groups according to the last digit of hospital number. Amniocentesis was obtained on all women at entry into the study, and labor was induced immediately if meconium staining was observed. If no meconium staining was present on initial amniocentesis, then subsequent monitoring was as follows: women in the amniocentesis group received weekly amniocentesis and were induced if meconium staining was present; and women in the CST group received an immediate CST, repeated weekly if normal. Labor was induced in significantly more women in the amniocentesis group than the CST group (11/90 [12 percent] vs. 29/90 [2 percent], respectively; $p < 0.005$). There were no statistically significant differences between testing groups for any outcome, including Apgar score < 7 at 1 minute, Apgar score < 7 at 5 minutes, low birthweight (< 10th percentile), neonatal morbidity, perinatal death, cesarean sections, or abnormal labor (prolonged latent phase, primary dysfunctional labor, secondary arrest of dilatation, or arrest). However, the proportion of babies with Apgar scores less than 7 at 1 and 5 minutes was two-fold higher in the amniocentesis group; the study may have been underpowered to detect this difference.

A single observational study (Devoe and Sholl, 1983) correlated CST results with the clinical outcomes of fetal distress and low Apgar score at 5 minutes (Table 10). Seventy-two of 248 women had labor induced either electively (n = 39) or for abnormal test results (n = 33). Twenty-two women had nonreactive NST followed by positive CST, and 17 women had nonreactive NST but negative CST. The positive predictive value of the CST component of the sequential testing strategy (NST followed by CST if NST is nonreactive) was poor for prediction of low Apgar scores or fetal distress.

In summary, CST is at least equivalent to amniocentesis for meconium staining in terms of outcomes, with significantly fewer inductions; perhaps on the basis of this trial, amniocentesis is no longer used for this indication. In the setting of prolonged pregnancy, CST, when used sequentially for followup of abnormal NST, has good negative predictive value but poor positive predictive value, based on one observational study.

CST using nipple stimulation. We did not identify any studies where nipple stimulation was the sole method for performing contraction stress tests in the management of prolonged pregnancy.

Amniotic fluid measurements. We identified one relevant randomized trial. Alfirevic, et al., compared two ultrasonographic measurements of oligohydramnios, namely amniotic fluid index (AFI) < 7.3 and maximum pool depth (MPD) < 2.1 cm, among 500 women at greater than 40 weeks gestation (Alfirevic, Luckas, Walkinshaw, et al., 1997). Both groups also had NST every

3 days. There were no differences in fetal outcomes between the two strategies; however, abnormal NST was more often an indication for induction in the AFI group than in the MPD group (15 percent vs. 8 percent; $p = 0.04$). The overall rates of induction of labor were not statistically different between groups (87/250 vs. 77/250; $p = 0.39$). There was a trend toward cesarean section for fetal distress being more common in the AFI group than in the MPD group (8 percent vs. 4 percent; $p = 0.09$). One possible explanation for this is a lower threshold for a diagnosis of fetal distress or for performing cesarean section in the presence of nonreassuring fetal heart rate tracings or abnormal antepartum NST results. Since such results were more common in the AFI group, it is not surprising that cesareans for fetal distress also were more common.

In a comparative cohort study, Eden, et al., reported a series of 585 patients managed in one of three ways (based on temporal changes in the protocol used): (1) weekly NST with CST for nonreactive NST (from November 1, 1978 through August 31, 1979); (2) semi-weekly NST with biophysical profile for nonreactive NST (from September 1, 1979 through December 31, 1980); or (3) semi-weekly NST with biophysical profile for nonreactive NST, plus weekly AFV measurement (from January 1, 1981 through August 31, 1981) (Eden, Gergely, Schifrin, et al., 1982). The groups employing the biophysical profile had lower incidences of low Apgar score at 5 minutes, meconium aspiration, stillbirth, fetal distress requiring intervention (persistent abnormal FHR patterns), and morbidity (defined as presence of any of following: fetal distress requiring intervention, 5-minute Apgar score < 7, neonatal resuscitation, postmaturity syndrome, or meconium aspiration). However, the rate of cesarean sections was significantly higher in the groups using the biophysical profile than in the group using NST + CST alone (NST + CST, 11.5 percent; NST + biophysical profile, 29.9 percent; NST + AFV + biophysical profile, 29.4 percent; 1 vs. 2, $p < 0.05$; 1 vs. 3, $p < 0.05$). This suggests that tests using the biophysical profile may be more sensitive at identifying fetuses at risk, but that subsequent induction resulted in higher cesarean section rates. Alternatively, as discussed above, physician thresholds for performing cesarean section may be quite different based on knowledge of antepartum test results. Despite the higher rates of cesarean section, the incidence of fetal distress requiring intervention was substantially lower in the groups using biophysical profile testing in addition to NST (NST + CST, 21.8 percent; NST + biophysical profile, 4.5 percent; NST + AFV + biophysical profile, 5.5 percent; 1 vs. 2, $p < 0.05$; 1 vs. 3, $p < 0.05$).

Tongsong and Srisomboon (1993) performed NST and AFV in 242 women at 42 weeks or more in gestational age. AFV was more accurate than NST in predicting intrapartum fetal distress ($p < 0.05$) (AFV: sensitivity, 73 percent; specificity, 91 percent; positive predictive value, 27 percent; negative predictive value, 99 percent; NST: sensitivity, 64 percent; specificity, 82 percent; positive predictive value, 14 percent; negative predictive value, 98 percent). Given that the definition of intrapartum fetal distress included moderate to severe variable decelerations, which would be more likely in a setting of oligohydramnios, which in turn would be more likely to be detected with ultrasound, these results are not surprising.

Table 11 summarizes sensitivity, specificity, and positive and negative predictive values for predicting reported perinatal and maternal outcomes, using amniotic fluid measurement with various criteria for abnormality. In general, specificity is markedly better than sensitivity, while negative predictive value is better than positive predictive value, as was also the case with NST and CST.

Abdominal palpation. As part of an investigation of the value of ultrasound evaluation of amniotic fluid volume in predicting adverse outcomes, Crowley, et al., also evaluated the performance of clinical assessment of AFV by abdominal palpation. This technique had a false positive rate of 25 percent and a false negative rate of 43 percent for predicting "significant meconium staining or absent amniotic fluid" at the time of amniotomy (Crowley, O'Herlihy, and Boylan, 1984).

Simple biophysical profile. Table 12 describes the individual components of the various biophysical profiles employed in the studies included in this report. One randomized trial and four noncomparative studies provide data on a simple biophysical profile (NST plus measurement of amniotic fluid volume). The randomized trial compared a simple biophysical profile (NST + maximum pool depth [MPD]) with a complex biophysical profile consisting of NST, amniotic fluid index (AFI), fetal breathing movements, fetal tone, and fetal gross body measurements for antenatal monitoring (Alfirevic and Walkinshaw, 1995). There were more abnormal test results with the complex biophysical profile (47 percent vs. 21 percent; $p = 0.0013$), more inductions of labor (60 percent vs. 41 percent; $p = 0.04$), and more inductions associated with abnormal testing (39 percent vs. 15 percent; $p = 0.002$). There were no significant differences in clinical fetal or maternal outcomes. Cesarean section rates were nonsignificantly higher in the complex monitoring group (18 percent vs. 10 percent; $p = 0.22$).

Four studies described the accuracy of simple biophysical profiles for predicting a variety of outcomes (Arias, 1987; Bochner, Medearis, Ross, et al., 1987; Bochner, Williams, Castro, et al., 1988; Brar, Horenstein, Medearis, et al., 1989) (Table 13). Although Bochner, et al. (1987) reported high values for sensitivity and specificity of the simple biophysical profile for predicting low Apgar scores at 5 minutes and cesarean section for fetal distress, the confidence intervals around those estimates were wide because the 2-by-2 tables were based on a relatively small subset ($n = 62$) of the study's 845 patients. The other studies show relatively poor sensitivity and specificity.

Table 13 summarizes the results of studies of simple biophysical profiles. Again, in general, specificity for the various outcomes is better than sensitivity, while negative predictive value is consistently higher than positive predictive value.

Complex biophysical profile score. The randomized trial of Alfirevic and Walkinshaw (1995) comparing simple with complex biophysical profiles is discussed above. Three other studies reported data on the performance of a complex biophysical score (Table 14). Since the definition of "complex" varied between studies, the items used to calculate the scores in individual studies are shown in Table 12.

Arabin, Snyjders, Mohnhaupt, et al. (1993) compared the predictive ability of a biophysical profile consisting of NST, amniotic fluid assessment, fetal tone, fetal movements, and fetal breathing to a novel fetal assessment score consisting of five components: FHR pattern, uterine artery resistance by Doppler ultrasound, carotid artery resistance index by Doppler ultrasound, fetal tone (movements) by ultrasound, and fetal reflexes (magnitude and speed of movements) by ultrasound. In receiver operating characteristic (ROC) analysis, the fetal assessment score provided better prediction of fetal distress and low Apgar score at 1 minute than did the biophysical profile ($p < 0.001$) but not better prediction of low umbilical artery pH. Qualitatively, the difference was greatest for prediction of fetal distress, with less difference noted for prediction of low Apgar scores and none for prediction of low pH. This suggests that the fetal prediction score is better at discriminating results that correlate directly with its component tests (such as fetal distress defined by abnormal fetal heart rate patterns) than at true physiological measures of fetal compromise. One possible explanation for this could be interpretation of intrapartum fetal monitoring based on prior knowledge of antepartum test results.

Hann, et al., reported the results of biophysical profile monitoring in 131 women at 41 completed weeks gestation (Hann, McArdle, and Sachs, 1987). Positive predictive values for "poor neonatal outcome" (neonatal distress requiring admission to the neonatal intensive care unit, endotracheal intubation, use of positive pressure ventilation for more than 6 hours, and/or persistent fetal circulation) for the composite biophysical profile at a threshold of ≤ 6 was 14 percent; for individual components, positive predictive values were as follows: AFV, 17 percent; placental grading, 4 percent; fetal breathing movements, 5 percent; fetal tone/movements, 40 percent; and nonreactive NST, 14 percent. Negative predictive value for the composite biophysical profile was 94 percent; for individual components: AFV, 95 percent; placental grading, 91 percent; fetal breathing movements, 94 percent; fetal tone/movements, 95 percent; and reactive NST, 94 percent.

Gilson, O'Brien, Vera, et al. (1988) describe the association between twice weekly biophysical profile monitoring and low Apgar scores, fetal distress, and cesarean section for fetal distress among 178 women at greater than 42 weeks gestation. At the cut-point used (a score of 8), the test showed poor sensitivity across all outcomes, ranging from 0.08 to 0.27.

Table 14 summarizes the test characteristics reported in these studies. Again, specificity is generally better than sensitivity, while negative predictive value is consistently much higher than positive predictive value.

Doppler measurements of umbilical blood flow. Two studies reported data on the predictive value of Doppler measurements of umbilical artery blood flow (Battaglia, Larocca, Lanzani, et al., 1991; Farmakides, Schulman, Winter, et al., 1988) (Table 15). Battaglia, et al., evaluated Doppler velocimetry of umbilical artery used as screening test for predictive value in a case series (Battaglia, Larocca, Lanzani, et al., 1991). This was performed as a battery of tests including NST; amnioscopy; AFV; Doppler velocimetry of the uterine, umbilical, descending thoracic aorta, renal, and middle cerebral arteries; and a series of maternal blood measurements, including hPL, estriol, hematocrit, platelets, mean platelet volume, and uric acid. The criteria for decisionmaking about induction and delivery were not described. Doppler velocimetry was strongly associated with adverse outcomes, including "poor condition" (both 1- and 5-minute Apgar scores < 7 or infant admitted to NICU for asphyxia and/or meconium aspiration syndrome), oligohydramnios (largest pocket < 2 cm), meconium staining, and cesarean sections for fetal distress. Of note, 4 of 16 of these infants had birthweights greater than 4,000 grams; it is unclear to what extent these infants, who presumably had normal uteroplacental function, affected the results.

Farmakides, et al., reported on 140 high-risk pregnancies (33 percent were postdate) that were followed with NST and Doppler velocimetry (Farmakides, Schulman, Winter, et al., 1988). "Most" of the cases of fetal distress and cesareans for fetal distress came from the postdate subgroup. Nonreactive NST was significantly more sensitive at predicting cesarean section for fetal distress than Doppler. Since management decisions were based on NST results, this again raises the possibility of biased decisionmaking based on prior knowledge of antepartum test results.

Table 15 summarizes the results of these studies of Doppler. Again, negative predictive value is consistently higher than positive predictive value, although sensitivity appears to be improved relative to specificity compared with the other tests reviewed in this report.

Summary of tests to evaluate risks to the fetus associated with uteroplacental insufficiency. There are no randomized trials comparing antepartum testing by any method to no testing in women with prolonged pregnancy only. Data from one relatively large retrospective cohort (Bochner, Williams, Castro, et al., 1988) suggest an increased risk of adverse outcomes to the fetus, although confounding cannot be eliminated as a possibility for this observed association. Evidence from large registries shows consistently elevated risks of antepartum stillbirth with increasing gestational age, even in health systems where testing is available (see the section on "Risk of Perinatal Mortality" in chapter 1). Given this elevated risk, it is highly unlikely that a randomized trial of testing versus no testing could be performed in the United States without, at the least, extreme difficulty with recruitment. The low absolute risk of stillbirth makes sample size requirements prohibitive as well. For example, the estimated perinatal mortality at 41 weeks in terms of deaths per 1,000 ongoing pregnancies is approximately 1.2. A randomized trial would need over 40,000 women in each arm to determine a two-fold difference in risk of stillbirth between two competing methods of antepartum surveillance.

Because of the numerous methodological issues involved in evaluating specific antepartum tests (see discussion below), we are unable to conclude that any test or combination of tests is clearly superior to another. Only one randomized trial directly compared a more complex test with a simpler test (Alfirevic and Walkinshaw, 1995); this trial showed that the more complex test resulted in more interventions with no difference in outcomes. As with most tests, there appear to be consistent tradeoffs between sensitivity and specificity–tests that are more sensitive are likely to be less specific. We did not identify published data on inter- or intraobserver variability of these tests in the specific context of monitoring prolonged pregnancy or on the medical and nonmedical costs associated with specific tests and testing regimens.

We did find that, qualitatively, specificity for most tests was considerably better than sensitivity, while negative predictive value also was considerably better than positive predictive value. This means that women with "normal" test results are highly unlikely to experience the adverse outcomes used to determine a true "positive" test result. The high specificities reported may reflect biases in study design–when outcomes are either directly related to test results (such as nonreassuring fetal heart rate tracings after abnormal antepartum NST) or likely to be influenced by knowledge about the test results (such as cesarean section for fetal distress), specificity is likely to be relatively high.

This pattern of high negative predictive value in the setting of relatively low sensitivities has interesting implications for future management strategies. By Bayes' Theorem, positive predictive value can be expressed as:

True Positives/(True Positives + False Positives), or

[(Prevalence)*(Sensitivity)] /{[(Prevalence)*(Sensitivity)] + [(1-Prevalence)*(1-Specificity)]},

while negative predictive value is expressed as:

True Negatives/(True Negative + False Negatives), or

[(1-Prevalence)*(Specificity)] /{[(1-Prevalence)*(Specificity)] + [(Prevalence)*(1-Sensitivity)]}.

In practice, this means that increasing test sensitivity results in a higher negative predictive value, since the false negative rate decreases. Increasing test specificity results in a higher positive predictive value, since false positives decrease. Given the consistent pattern observed for all of the reviewed antepartum tests that specificity is higher than sensitivity, one would expect that positive predictive value would be higher than negative predictive value. The fact that the pattern is consistently the opposite suggests that it is the relatively low prior probability of adverse outcomes, the “prevalence” in the equations above, that drives the predictive values.

If this is the case, then the following points need to be considered:

- ♦ The main purpose of antepartum testing is primarily to avoid unexplained stillbirths and secondarily to avoid perinatal morbidity. In order to accomplish these things, tests with high negative predictive values are needed. One way to achieve this would be to improve the sensitivity of currently used antepartum testing technologies. Since it is unlikely that sensitivity can be increased without a subsequent decrease in specificity, this means that the positive predictive value of these tests will decrease further.

- ♦ If, as the reviewed studies suggest, the probability of adverse outcomes is currently what determines predictive values, then this means that the positive predictive value of antepartum testing will improve and the negative predictive value decline as gestational age increases, since the risk of stillbirth and other adverse events increases with gestational age. This proposition is dependent on the assumptions that (1) sensitivity and specificity are independent of gestational age, and (2) the outcomes reported in these studies are reasonable surrogates for stillbirth risk. This proposition is consistent with the data reported by Bochner, Williams, Castro, et al. (1988), according to which the positive predictive value for all adverse outcomes was better when testing began at 42 weeks (21.1 percent vs. 11.9 percent when testing began at 41 weeks), but the negative predictive value was worse (98.5 percent at 42 weeks vs. 99.1 percent at 41 weeks).

- ♦ Assuming that induction of labor does not carry increased perinatal risks compared with spontaneous labor, planned induction of labor at a given gestational age will always result in fewer expected adverse perinatal outcomes compared with testing strategies, since the negative predictive value of the tests will continue to decline as gestational age advances. At earlier gestational ages, where the risk is very low, the number of patients required to demonstrate this would be quite large.

These implications will be discussed further in the context of the trials of induction versus testing (Question 2).

Assessment of Risks to the Fetus and Mother Associated with Fetal Macrosomia

Because both mother and infant are at risk of injury secondary to macrosomia, various methods for estimating fetal weight have been evaluated. Macrosomia is usually defined as a newborn weight of greater than 4,000 grams or 4,500 grams; the clinical significance of birthweights between 4,000 and 4,500 grams is unclear, since risk of shoulder dystocia is greatest for infants over 4,500 grams (ACOG, 2000).

Clinical exam. Chauhan, et al., compared estimates of fetal weight by clinicians using Leopold maneuvers in early labor, sonographic measurements obtained by the same clinicians, and actual birthweight (Chauhan, Sullivan, Magann, et al., 1994). Clinical estimation was significantly more accurate than ultrasound estimation as measured by mean absolute error compared with actual weight (clinical, 322 ± 253 g; sonographic, 547 ± 425 g; $p < 0.001$), mean percentage absolute error (clinical, 8.9 ± 7.1 g/kg; sonographic, 14.8 ± 11.0 g/kg; $p < 0.001$), and percentage of estimates within 10 percent of actual birthweight (clinical, 65.4 percent; sonographic, 42.8 percent; $p < 0.005$).

The same group also compared maternal estimations by women with prior childbearing experience with clinical estimation (Chauhan, Sullivan, Lutton, et al., 1995). There were no significant differences in the accuracy of maternal estimates compared with clinical estimates.

Ultrasound. Chauhan, et al. (Chauhan, Sullivan, Magann, et al., 1994) found that clinical estimation was more accurate than ultrasonographic estimation by the same clinician (see above). Ultrasound was slightly more sensitive at predicting birthweight greater than 4,000 grams (55 percent vs. 50 percent, based on 20 cases).

Chervenak, et al., compared 317 women followed for prolonged pregnancy with twice weekly NST and AFT with100 control patients delivered between 38 and 40 weeks (Chervenak, Divon, Hirsch, et al., 1989). Fetal weights were also obtained, although it is unclear how often these measurements were performed. Overall incidence of birthweight greater than 4,000 grams was significantly higher in postdate patients (24 percent vs. 4 percent; $p < 0.05$), and cesarean section rates for arrest or protraction disorders were significantly higher when infants weighed more than 4,000 grams (22 percent vs. 10 percent; $p < 0.01$). Sensitivity of ultrasound for predicting birthweight greater than 4,000 grams was 61 percent, specificity 91 percent, positive predictive value 70 percent, and negative predictive value 87 percent. Morbidity associated with macrosomia was not reported. It is unclear to what extent clinicians managing the patients had access to the ultrasound reports. Since clinicians might have a lower threshold for diagnosing an arrest or protraction disorder in the setting of suspected macrosomia, this would result in a bias in favor of improved positive predictive value for ultrasound.

Gilby, et al., constructed ROC curves for the performance of two abdominal circumference cut-points (35 cm and 38 cm) for predicting macrosomia at two thresholds, 4,000 grams and 4,500 grams, from a series of 1,996 subjects who had ultrasounds within 7 days of delivery (Gilby, Williams, and Spellacy, 2000). At a cut-point of 35 cm, sensitivity for prediction of birthweight of 4,500 grams was 98.5 percent, specificity 64.6 percent, positive predictive value 9.1 percent, and negative predictive value 99.9 percent. At a cut-point of 38 cm, sensitivity was 53.6 percent, specificity 96.8 percent, positive predictive value 37.3 percent, and negative predictive value 98.3 percent. Morbidity associated with macrosomia was not reported. Whether these predictive values would be applicable in a different population is unclear.

O'Reilly-Green and Divon (1997) constructed ROC curves for ultrasonographic estimates of fetal weight, with an adjustment of 12.7 grams added to the estimated fetal weight (EFW) for each day elapsed between sonographic measurements and delivery. Areas under the ROC curve for prediction of birthweight greater than 4,000 grams were 0.85 and 0.93 to 0.95 for prediction of birthweight greater than 4,500 grams, indicating good discriminative ability. Relatively small relative increments in EFW had large impacts on sensitivity and specificity: for prediction of actual birthweight of greater than 4,000 grams, an EFW of 3,711 grams had a sensitivity of 85 percent and specificity of 72 percent, while an EFW of 4,000 grams had a sensitivity of 56 percent and a specificity of 91 percent. For prediction of birthweight greater than 4,500 grams, an EFW of 4,192 grams had sensitivity of 83 percent and specificity of 92 percent, while an EFW of 4,500 grams had a sensitivity of 22 percent and a specificity of 99 percent. Again, no correlation with outcomes associated with fetal macrosomia were reported.

Test performance characteristics for studies reporting association between estimated fetal weight and macrosomia are shown in Table 16.

Summary: Tests for predicting fetal macrosomia. There is a clear tradeoff between sensitivity and specificity of markers for estimating fetal weight. The definition of macrosomia also plays a role. In studies in women with prolonged pregnancy, sensitivities for detection of birthweight greater than 4,000 grams range from 56-89 percent, with specificities of 72-93 percent; positive predictive values at this threshold range from 49-93 percent, with negative predictive values of 87-94 percent. At a threshold of 4,500 grams, sensitivity ranges from 14-99 percent and specificity from 65-99 percent, with positive predictive values of 9-44 percent and negative predictive values of 96-100 percent. Positive predictive value at the more clinically significant 4,500 gram threshold is worse than at 4,000 grams (not surprisingly, since the probability of a weight greater than 4,500 grams is much lower than for 4,000 grams). However, translation of even this diagnostic test accuracy into clinical strategies that significantly reduce injury risk to either mother or infant at an acceptable cost in terms of iatrogenic complications or resource use is difficult.

Prior suspicion of fetal macrosomia does not appear to result in improved outcomes for either mother or infant. Weeks, et al., reported a retrospective series of 504 infants with birthweight greater or equal to 4,200 grams (Weeks, Pitman, and Spinnato, 1995). In 102 patients, macrosomia was suspected, while it was not in the remaining 402. Cesarean delivery rates were significantly higher in the suspected group (52 percent) compared with the unsuspected group (30 percent), a difference attributable to a higher rate of labor induction and failed induction. Among patients undergoing vaginal delivery, shoulder dystocia occurred in 24.5 percent of the predicted group and 16.7 percent in the not predicted group, a difference that was not statistically significant (which may be due to lack of power).

Even better evidence of a lack of benefit comes from a trial in which women at 38 weeks or more with estimated birthweights between 4,000 and 4,500 grams based on ultrasound were randomized to either immediate induction or expectant management. There were no statistically significant differences in cesarean delivery rate, instrumental delivery rate, or incidence of shoulder dystocia between the two groups (Gonen, Rosen, Dolfin, et al., 1997). There were trends toward higher instrumental delivery rates in induced nulliparous women (26.2 percent vs. 15 percent in expectantly managed nulliparous women) and higher cesarean section rates in expectantly managed multiparous women (16.2 percent vs. 10.9 percent in induced multiparous women). Other maternal outcomes, such as perineal or vaginal trauma, were not reported. The study was underpowered to detect differences in neonatal morbidity; overall rates were low (9/134 in the induction group and 11/139 in the expectant group), with six or fewer cases of any single type of morbidity (cephalohematoma, with nine cases, was most common).

Rouse, Owen, Goldenberg, et al., (1996) estimated based on available data that a policy of elective cesarean section for an estimated fetal weight of 4,500 grams or more would result in 3,695 cesarean deliveries at a cost of over $8 million to prevent one permanent brachial plexus injury.

In summary, methods for detection of macrosomia defined as birthweight greater than 4,500 grams are imprecise. There is evidence that clinical measurements, including multiparous patients' own estimates, are as accurate as ultrasound. Available data suggest that there is no benefit to mother or infant from induction of labor for suspected macrosomia (when defined as estimated weights between 4,000 and 4,500 grams). While an estimate of fetal weight in theory may have some benefit in management of labor (such as avoidance of operative vaginal deliveries in settings where shoulder dystocia risk is higher), available observational data suggest that suspicion of macrosomia prior to labor does not improve outcomes. There is no evidence that ultrasonographic measurement of fetal weight to detect macrosomia in the setting of prolonged pregnancy improves maternal or neonatal outcomes.

Assessment of the Likelihood of Successful Induction

Cervical examination (Bishop score). The Bishop score was first reported in 1964 as a predictor of the likelihood of a successful induction (Bishop, 1964). The score is based on five components: cervical dilation, cervical effacement, cervical consistency, cervical position, and fetal station (Table 17).

In Bishop's original report (Bishop, 1964), induction was successful in 100 percent of cases (no denominator given) when the Bishop score was greater than 9. Data for lower scores were not given, and notably, all inductions were apparently in multiparous patients, since "[o]wing to the unpredictability of the duration of labor in the nullipara, even in the presence of apparently favorable circumstances, induction of labor brings little advantage for either obstetrician or patient." There was a statistically significant negative correlation between score and interval from examination to spontaneous delivery, but confidence intervals were quite wide (quantitative data were not provided, only a graphic representation).

Three studies provided limited data on the predictive value of Bishop scores (Harris, Huddleston, Sutliff, et al., 1983; Mouw, Egberts, Kragt, et al., 1998; Witter and Weitz, 1989). Harris, et al., reported that dilatation, effacement, and station were more predictive of interval between examination and spontaneous delivery in prolonged pregnancy than consistency and position (Harris, Huddleston, Sutliff, et al., 1983). Witter and Weitz (1989) found that Bishop scores at baseline in women induced at 42 weeks were statistically significantly lower in women who underwent cesarean delivery than in those with vaginal delivery, but that the absolute difference was small; significant overlap made the test a poor discriminator of successful induction (Table 18). Mouw, et al., reported that a Bishop score greater than 5 at 41 weeks had sensitivity 0.67 (95 percent CI, 0.48 to 0.82) and specificity 0.77 (95 percent CI, 0.54 to 0.92) for predicting birth within 3 days; however, only 74 percent of patients in this study had Bishop scores recorded (Mouw, Egberts, Kragt, et al., 1998).

The relatively poor discrimination of the Bishop score in predicting either labor or subsequent successful induction in prolonged pregnancy is magnified by the inherent unreliability of many of its component measures. Significant interobserver variability has been reported in measurement of cervical effacement (Goldberg, Newman, and Rust, 1997; Holcomb and Smeltzer, 1991). Furthermore, significant intra- and interobserver variability has been described for assessment of cervical dilatation (Phelps, Higby, Smyth, et al., 1995; Tuffnell, Bryce, Johnson, et al., 1989)

Fibronectin. Three studies were identified that evaluated the possible use of fetal fibronectin (fFN) obtained from cervicovaginal secretions, a sensitive marker for impending labor, in the management of prolonged pregnancies (Table 19). Tam, et al., measured fetal fibronectin in 58 women at term or beyond, scheduled for induction with PGE_2 suppositories (Tam, Tai, and Rogers, 1999). Thirty women were negative and 28 positive for fibronectin prior to the placement of the suppositories. There was a trend towards a higher gestational age in fibronectin-positive patients (median 294 days, range 280-294, compared with a median of 281 days, range 272-294, in negative patients). Median interval from induction to delivery was significantly lower in fibronectin-positive patients (760 minutes vs. 1,285 minutes). Fibronectin positivity was a reasonable predictor of vaginal delivery (sensitivity 36 percent; specificity 79 percent; positive predictive value 84 percent; negative predictive value 28 percent). Results in this study were not stratified by gestational age or by indication for induction.

Mouw, et al., measured fetal fibronectin at 41 weeks (Mouw, Egberts, Kragt, et al., 1998). A positive fFN test (≥ 50 ng/ml) had sensitivity of 0.71 (95 percent CI, 0.58 to 0.86) and specificity of 0.64 (95 percent CI, 0.48 to 0.78) for predicting birth within 3 days. The change from negative to positive fFN values often occurred between 1 and 4 days before birth in women with a spontaneous onset of labor. The mean interval between positive test and birth was 2.5 ± 2.5 days (range, 0-11).

Imai and colleagues measured vaginal fFN and a panel of cytokines (interleukin 1-beta, interleukin-6, interleukin-8, and tumor necrosis factor alpha) weekly in 122 women from 36 through 42 weeks (Imai, Tani, Saito, et al., 2001). Vaginal fFN was inversely correlated with sampling to delivery interval (r = -0.40). At a threshold of > 50 ng/ml, fFN had a sensitivity of 90 percent, a specificity of 50 percent, a positive predictive value of 75 percent, and a negative predictive value of 75 percent for predicting delivery within 7 days. Interleukin 1-beta was the only cytokine with reasonable performance, but it was less able to discriminate than fFN (sensitivity 55 percent, specificity 76 percent). Results were not stratified by parity or gestational age.

Summary: Tests for assessing the likelihood of successful induction. The Bishop score has a long history in obstetric decisionmaking. Clearly, clinically detectable changes in the cervix take place prior to the onset of labor, and the likelihood of a successful induction should be greater the closer a given patient is to spontaneous labor. However, the documented substantial inter- and intraobserver variability in the components of the Bishop score suggest that its ability to discriminate between women likely to have a successful induction of labor and those unlikely to have a successful induction may be relatively poor. Certainly, given this inherent variability and the discrete nature of its components, changes in the global Bishop score are less than satisfactory primary outcomes for studies of induction or cervical ripening agents.

Data on the clinical utility of fetal fibronectin as a decisionmaking tool in managing prolonged pregnancy are insufficient to draw conclusions. Fetal fibronectin may have potential as a tool for helping to identify women likely to deliver spontaneously within the next 7 days, which in turn may help guide decisionmaking about antepartum testing versus induction.

Methodological Issues

Study Design

- Choice of appropriate outcome measures: Many of the most important outcome measures, especially stillbirth, are so rare that studies using these outcomes are almost impossible to perform. Surrogate markers therefore are not inappropriate, but their clinical relevance is not always clear. For example, although meconium aspiration is a significant adverse outcome with potential for long-term negative sequelae, the presence of meconium-stained amniotic fluid alone is not. Intrapartum abnormal fetal heart rate tracings themselves are subject to significant observer variability (Ayres-de-Campos, Bernardes, Costa-Pereira, et al., 1999; Bernardes, Costa-Pereira, Ayres-de-Campos, et al., 1997; Donker, van Geijn, and Hasman, 1993; Lidegaard, Bottcher, and Weber, 1992), and interpretation may be influenced by prior knowledge of antepartum test results, making fetal heart rate patterns, or cesarean section decisions based on these patterns, less than ideal as surrogate markers of fetal compromise.

- Bias: Many of the studies reviewed either did not state whether clinicians managing patients were aware of test results or definitely stated that these results were available. Since knowledge of these results could affect both interpretation of outcomes (as discussed above) or thresholds for decisionmaking (e.g., greater reluctance to use oxytocin to augment labor if prior antepartum testing was abnormal, or a lower cesarean section threshold for arrest of dilatation or descent if macrosomia were suspected), the ability of tests to predict these outcomes could be falsely elevated.

- Resource use: Data on the medical and nonmedical costs of any of the tests reviewed are lacking.

Statistical Issues

- Inappropriate summary measures and tests: Many studies used means or t-tests for variables such as Bishop scores, Apgar scores, or parity, where values other than integers are meaningless.

- Sample size: Few studies discussed sample size issues.

- Failure to account for variability: No study attempted to account for the effects of observer variation on the precision of estimates. For tests where quantitative values are used to establish a threshold for normal and abnormal, this variability will have implications for the precision of sensitivity and specificity.

Summary

- The risk of antepartum stillbirth clearly increases with increasing gestational age. Although definitive evidence that antepartum testing at some point after 40 weeks reduces perinatal mortality is not available, there are some data consistent with an increased risk of adverse outcomes in women who do not get tested (Bochner, Williams, Castro, et al., 1988; Fleischer, Schulman, Farmakides, et al., 1985). The most appropriate time to begin antepartum testing in otherwise low-risk women is unclear. An excellent decision analysis of antepartum testing in high-risk women prior to 40 weeks illustrated that the tradeoffs are between the risk of stillbirth, the risk of neonatal death, and the sensitivity and specificity of the test (Rouse, Owen, Goldenberg, et al., 1996). Since the risk of neonatal death in an otherwise uncomplicated pregnancy at term is quite low, the main issues are the stillbirth risk and test characteristics. Unfortunately, our review does not allow precise estimation of the test characteristics of any of these tests in detecting infants at greatest risk for stillbirth in otherwise uncomplicated pregnancies after term.

- As the sensitivity of antepartum testing for predicting surrogate markers of fetal compromise increases, specificity decreases. Testing strategies involving a combination of fetal heart rate monitoring and ultrasonographic measurement of amniotic fluid volume appear to have the highest levels of sensitivity; however, methodological issues and variability in specific tests and testing strategies prohibit definitive conclusions about which test or combination of tests has the best performance.

- Qualitatively, we found that specificity was much higher than sensitivity for most of the outcomes measured, but negative predictive values were much higher than positive predictive values, suggesting that outcome probability is currently the most important determinant of test performance. This in turn implies that the negative predictive value will decrease as gestational age advances, and rates of adverse outcomes due to false negative test results will increase, if sensitivity and specificity of antepartum tests are independent of gestational age. Identifying the most appropriate time to begin testing (or to consider induction) is ultimately dependent on identifying threshold risks of adverse outcomes when weighed against the risks and costs of intervention. We did not identify any data that would allow estimation of that threshold risk.

- Low positive predictive values mean that intervention rates will be relatively high. The degree to which individual women, or society, are willing to trade off risk of adverse fetal outcomes due to prolonged pregnancy, versus the potential for iatrogenic adverse outcomes associated with interventions, is unclear. How variability in the value women place on the nature of the process of labor and delivery (minimal intervention vs. use of the full range of available obstetric, anesthetic, and pediatric technologies) factors into decisionmaking is also unclear.

- Clinical assessment is equivalent to ultrasound in predicting macrosomia. However, there is no evidence that prior knowledge of estimated fetal weight improves outcomes for either infant or mother.

- Clinical examination of the cervix may help predict successful induction. However, individual components of the examination exhibit substantial inter- and intraobserver variability.

- Published data do not allow estimation of the cost-effectiveness of tests of fetal wellbeing.

Question 2: What is the direct evidence comparing the benefits, risks, and costs of planned induction versus expectant management at various gestational ages?

Approach

As with all of the questions addressed in this report, the issue of the appropriate gestational age to consider " postdate" or "postterm" was difficult to resolve. After extensive discussion with the project's advisory panel, a consensus was reached that we would include any articles where the proposed benefit of the planned induction was reduction in maternal or fetal risk associated with prolonged pregnancy, even at 40 weeks gestation. Active interventions performed prior to or shortly after term (such as nipple stimulation or membrane sweeping) that are designed to decrease the proportion of women who go beyond 41 or 42 weeks are discussed under Question 3, below.

Up to this point in the report, we have:

- Found evidence from observational studies of an increasing risk of adverse perinatal events as gestational age advances beyond term. Although the precise degree of this risk is unclear and may be affected by confounding, the pattern is quite consistent.

- Found in our review of antepartum tests of fetal well being in prolonged pregnancy that the sensitivity of such tests was much lower than the specificity, while the negative predictive value was much higher than the positive predictive value.

- Discussed the fact that these two findings, when taken together, suggest that the negative predictive value of antepartum testing will decrease as gestational age advances.

If negative predictive value does decrease with advancing gestational age, then elective induction has the potential to improve outcomes by preventing adverse perinatal outcomes due to false negative test results. Whether this is the case, and whether elective induction is associated with an excess of other adverse maternal outcomes compared with expectant management and testing, is the focus of this section of the report.

Throughout this section, we use the term "expectant management," as defined by the authors of the studies reviewed, to refer to some form of ongoing assessment of fetal well being, with induction of labor based on the results of testing or upon reaching a specified gestational age in accordance with a predefined set of guidelines. As stated above, we did not identify any randomized trials that provided data on the specific population of interest where no intervention (induction or testing) was performed.

As with studies of testing, the outcomes assessed in these trials were quite variable. All studies reported on perinatal mortality and cesarean section rates, in some cases stratified by indication for induction (elective or based on abnormal test results). Additional markers of perinatal or maternal morbidity—including Apgar scores at 1 and 5 minutes, umbilical arterial pH, the presence of meconium-stained amniotic fluid, abnormal fetal heart rate tracings during labor, instrumental deliveries, diagnosis of meconium aspiration, and admissions to neonatal intensive care units—were inconsistently reported.

None of the included trials was able to blind physicians, midwives, and nurses to the allocated intervention or to the results of antepartum testing. Because of this, outcomes that are dependent on interpretation of fetal monitoring (such as the proportion of cesarean sections performed for fetal distress, or the overall incidence of abnormal fetal heart rate tracings) are unreliable. A diagnosis of fetal distress may be more likely in the setting of an induction performed in the expectant management arm after abnormal antepartum monitoring. Even with a normal intrapartum tracing, thresholds for performing cesarean section or operative vaginal delivery in the setting of prolonged second or third stages of labor might be different if the provider is aware of previous abnormal antepartum tests. Because of these difficulties, we focus on the overall cesarean section rate and neonatal outcomes less susceptible to bias, such as the Apgar score, pH, and admissions to the neonatal intensive care unit. Even these immediate outcomes do not provide information on the impact of maternal interventions on longer-term health outcomes of these children.

Results

Trials Identified

The literature search identified 17 relevant publications reporting on 15 separate trials (see Evidence Table 2). In two cases, initial trial reports were followed by publications describing further analyses conducted on the same populations: Pearce and Cardozo (1988) reported the results of supplementary analyses conducted on the population first described by Cardozo, Fysh, and Pearce (1986), and Goeree, Hannah, and Hewson (1995) reported the results of a cost-effectiveness analysis of data collected during the Canadian Multicenter Post-term Pregnancy Trial (Hannah, Hannah, Hellmann, et al., 1992).

The included trials were published between 1983 and 1997. The number of subjects in each trial was fairly small, except for the Canadian trial (Hannah, Hannah, Hellmann, et al., 1992). The overall median number of subjects was 200, ranging from 22 (Martin, Sessums, Howard, et al., 1989) to 3,418 (Hannah, Hannah, Hellmann, et al., 1992).

Benefits

Effects on perinatal mortality. The included studies suggest that induction results in fewer perinatal deaths than does expectant management. Table 20 summarizes perinatal deaths not due to congenital abnormalities in the two management groups. There were a total of seven deaths in the monitoring group compared with no deaths in the induction group.

A meta-analysis performed as part of a recent Cochrane review (Crowley, 2000) showed that this reduction in perinatal mortality with induction is significant only at 41 weeks or later (summary odds ratio [OR], 0.13; 95 percent confidence interval [CI], 0.01 to 2.07 before 41 weeks vs. summary OR, 0.23; 95 percent CI, 0.06 to 0.90 at 41 weeks or later).

Effects on perinatal morbidity. Other perinatal outcomes examined included Apgar scores. Of the 15 included trials, 14 evaluated Apgar scores, and all but one of these found substantially equal scores in the induction and monitoring groups. Dyson, Miller, and Armstrong (1987) reported that a higher proportion of babies in the monitoring group had Apgar scores < 7 at 1 minute (21 percent vs. 11 percent in the induction group); however, similar proportions of infants in the two groups had scores < 7 at 5 minutes. There is evidence, based on these trials, to conclude that Apgar scores do not change significantly when comparing induction versus monitoring of pregnancies.

Potential maternal benefits. Only one trial (Cardozo, Fysh, and Pearce, 1986) measured patient satisfaction, patient preferences, or quality of life. There were no significant differences in the proportion of patients “pleased” with (49 percent, planned induction; 53 percent, expectant management) or “disappointed” by (15 percent, planned induction; 11 percent, expectant management) their management.

Risks

Perinatal morbidity and mortality. Hyperstimulation of the uterus from induction agents can result in fetal compromise, leading to the need for cesarean section or even fetal death. Because fetal compromise in labor with subsequent need for cesarean section is also associated with prolonged gestation, differences in “risks” for fetal compromise between planned induction and expectant management are the inverse of differences in “benefits” and are discussed above.

Continued fetal growth during expectant management could conceivably lead to an increased risk of macrosomia and shoulder dystocia. In the study by Dyson, Miller, and Armstrong (1987), the proportion of infants with a birthweight greater than 4,000 grams was higher in the expectant management group (28.2 percent) than in the induction group (19.1 percent), though the difference did not reach statistical significance, and no correlation with shoulder dystocia or birth injury was reported. Katz, Yemini, Lancet, et al. (1983) also reported that the incidence of birthweight greater than 4,000 grams was higher in the expectant management group (29.5 percent vs. 7.9 percent; $p < 0.05$), but again no correlation with birth injury was reported. Ohel, Rahav, Rothbart, et al. (1996) found no difference in the proportion of infants with a birthweight greater than 4,000 grams (8.6 percent vs. 8.7 percent). Augensen, Bergsjø, Eikeland, et al. (1987) reported only one case of “difficult shoulder delivery” in the entire study.

In the two large multicenter trials comparing planned induction and expectant management, there were no significant differences in reported rates of macrosomia, shoulder dystocia, or birth injury to the fetus. In the National Institute of Child Health and Human Development (NICHD) Maternal-Fetal Network Trial (National Institute of Child Health and Human Development Network of Maternal-Fetal Medicine Units, 1994), the incidence of birthweight greater than 4,500 grams was similar in the two induction arms and the expectant management arm, and there was only one case of nerve injury (in one of the induction arms). In the even larger Canadian Multicenter Post-term Pregnancy Trial (Hannah, Hannah, Hellmann, et al., 1992), neither the proportion of infants with a birthweight greater than 4,500 grams (4.6 percent in the induction group vs. 5.5 percent in the expectant management group), nor the incidence of shoulder dystocia (1.4 percent in the induction group vs. 1.6 percent in the expectant group) was significantly different in the two groups.

These results suggest, as would be expected, that continued growth occurs in most infants managed expectantly, resulting in higher proportions of infants over 4,000 grams. Since there is debate as to whether weights between 4,000 and 4,500 grams have any clinical relevance (ACOG, 2000), it is not surprising that there are no reported differences in birth injury. The fact that trials that defined macrosomia as greater than 4,500 grams found no difference in either the proportion of babies weighing more than 4,500 grams or incidence of shoulder dystocia suggests that elective induction at a predefined gestational age does not have prophylactic benefit—i.e., induction at a given gestational age prior to the development of “macrosomia” does not have an impact on shoulder dystocia.

Cesearean section. Of the 15 included trials, two found a statistically increased risk of overall cesarean section with induction, while three trials found a statistically increased risk of overall cesarean section with expectant monitoring (Table 21).

Meta-analysis and subgroup analyses performed as part of a recent Cochrane review (Crowley, 2000) found no significant differences in cesarean delivery rates in any group or subgroup (Table 22). If anything, cesarean rates tend to be slightly lower in the elective induction groups.

Hannah, et al., published an interesting reanalysis of the Canadian study in 1996 (Hannah, Huh, Hewson, et al., 1996). In this new analysis, women who were randomized to induction or expectant management were stratified based on whether labor was ultimately induced or spontaneous. In the induction arm, 772/1,149 women (67.7 percent) were induced, while 377/1,149 (33.3 percent) went into spontaneous labor prior to scheduled induction. In the expectant management group, 405/1,128 (35.9 percent) were induced for various indications, while 723/1,128 (64.1 percent) went into spontaneous labor. There were no significant differences in cesarean section rates between women randomized to induction who were induced (29.5 percent), women randomized to induction who went into spontaneous labor (25.7 percent), and women who were managed expectantly who went into spontaneous labor (25.7 percent). However, the cesarean section rate was significantly increased in women randomized to expectant management who were induced (42.0 percent). These women were significantly more likely to be nulliparous, to have a closed cervix at the onset of labor, and to have a longer interval from induction to delivery. When compared with the expectantly managed women in spontaneous labor, they had significantly higher cesarean section rates for fetal distress or dystocia; such differences were not seen when the two subgroups in the induction arm were compared.

These differences are consistent with several findings discussed earlier in this report:

- Women whose onset of labor is considerably later than average may represent a distinct subgroup with different physiological characteristics of the uterus and cervix. This is consistent with the higher proportion of women with closed cervices and may also explain the higher rates of cesarean section for dystocia. This also may be related to parity. Presumably, women are included in this group who reach a predefined date for induction without going into spontaneous labor and with normal antepartum testing.

- Provider knowledge of antepartum testing results may affect thresholds for cesarean delivery. It seems likely that providers caring for women whose inductions were indicated because of abnormal antepartum tests would be less tolerant of intrapartum fetal heart rate abnormalities or less likely to tolerate labor progress that was slower than average. This would explain some of the differential rates by indication.

- As Crowley (2000) points out, women induced in the expectant management arm were less likely to receive prostaglandins. This would be a bias in favor of induction. The reanalysis by

Hannah and colleagues (Hannah, Huh, Hewson, et al., 1996) models this based on assumptions about prostaglandin efficacy, and finds that, at worst, there would be no difference in cesarean section rates between groups. In addition, our review of the literature on induction agents (discussed under Question 3) suggests that the effectiveness of prostaglandins in terms of expediting delivery may be proportional to risk of fetal heart rate abnormalities in labor. If this is the case, then any decrease in cesarean section rates for failed induction or dystocia might well be accompanied by an increase in cesarean sections for fetal distress.

In summary, the randomized trial literature consistently shows that elective induction does not result in increased cesarean section rates compared with management strategies based on antepartum testing. If anything, cesarean section rates are slightly lower in women who are electively induced.

Operative vaginal delivery. No studies reported specifically on maternal trauma related to vaginal delivery. Because operative vaginal delivery is clearly associated with an increased risk of maternal injury (Johanson and Menon, 2001), evidence of a difference in the rates of operative vaginal delivery in one group or the other would be suggestive of an increased risk of trauma to the pelvic floor, vagina, or perineum. In seven of the eight studies where this outcome was reported (Bergsjø, Huang, Yu, et al., 1989; Cardozo, Fysh, and Pearce, 1986; Egarter, Kofler, Fitz, et al., 1989; El-Torkey and Grant, 1992; Hannah, Hannah, Hellmann, et al., 1992; Herabutya, Prasertsawat, Tongyai, et al., 1992; Martin, Sessums, Howard, et al., 1989), there were no significant differences between the induction and expectant management groups. In the remaining trial (Hedén, Ingemarsson, Ahlström, et al., 1991), there was a significant difference, with 2.8 percent of the induction group and 15.5 percent of the expectant management group undergoing operative vaginal delivery ($p < 0.01$); the majority of these deliveries in both groups were for “secondary arrest.” There are no obvious reasons why the results of this study varied so dramatically from the others. Mean birthweight in the two groups was similar. The standard deviation of the preintervention Bishop score was slightly wider in the expectant management group, and the method of randomization was based on a registration number rather than on randomly generated numbers. One possible explanation for the study’s finding on operative vaginal delivery is that the pseudorandomization scheme resulted in some systematic differences in the groups. Another possibility is that use of oxytocin for labor augmentation may have been less aggressive in the expectant management group for some reason.

Overall, the studies reviewed suggest that there is no difference in operative vaginal delivery rates between expectant management and planned induction protocols.

Other maternal risks. There were no differences in the risk of maternal infection or other morbidity in three of the four trials that reported these outcomes (El-Torkey and Grant, 1992; National Institute of Child Health and Human Development Network of Maternal-Fetal Medicine Units, 1994; Witter and Weitz, 1987). In the remaining, very small trial (Martin, Sessums, Howard, et al., 1989), the proportion of women with “maternal morbidity” was higher in the induction arm (4/12, or 33 percent) than in the expectant management arm (2/10, or 20 percent). No significance testing was reported.

Costs and Resource Use

Direct measures of cost. Only two studies reported direct measures of cost, the Canadian Multicenter Post-term Pregnancy Trial (Hannah, Hannah, Hellmann, et al., 1992) and a smaller study by Witter and Weitz (1987). The Canadian study found that induction of labor was associated with a lower cost compared with monitoring. The mean cost per patient (in 1991 Canadian dollars) of a prolonged pregnancy managed through monitoring was $3,132 (95 percent CI, $3,090 to $3,174), compared with induction, which cost $2,939 (95 percent CI, $2,898 to $2,981) per patient. The difference between the two groups ($193 per patient) was statistically significant. The authors of the study estimated that switching to planned induction could save up to $8 million per year in Canada.

Witter and Weitz (1987) found, on the contrary, that mean costs were higher for planned induction than for monitoring by approximately $250 per patient. This study had a much smaller patient population (n = 200). Because costs frequently are not normally distributed, the effects of a few patients with complications or very long stays may be magnified compared with a larger study.

Indirect measures of resource use. Several studies that did not report direct costs did report outcomes that are indirect measures of resource use, such as overall length of maternal or infant stay in the hospital. The extent to which these results are generalizable is limited, since length of stay varies internationally and has changed dramatically in the United States over recent years. Moreover, overall length of stay may not be entirely related to overall resource use (Tai-Seale, Rodwin, and Wedig, 1999). For women delivering in a hospital, the majority of resource use occurs during the time from admission to delivery, with a sharp decrease after delivery and even further decreases after the first 24 hours. Thus, even if the mean length of stay is equivalent between two groups, the resource use may vary widely depending on what proportion of the time was spent in the delivery suite. In addition, studies that report only hospital use and not outpatient use of resources (for antepartum testing, other office visits, etc.) will not reflect the overall medical costs of a particular strategy. Finally, none of the included studies addressed the nonmedical costs—such as transportation, time lost from work, child care for women with other children, and so on—associated with various strategies for managing prolonged pregnancy.

Table 23 shows reported mean maternal lengths of stay for the six trials where this was reported. There are no obvious trends. Because reporting of the proportion of time spent in labor versus postpartum was minimal, no additional inferences about relative resource use can be drawn.

Only one study (Dyson, Miller, and Armstrong, 1987) reported data on mean neonatal length of stay, with no significant differences between the induction and expectant management groups (3.0 days vs. 3.3 days, respectively).

Tables 24, 25, and 26 summarize perinatal and maternal outcomes and resource use for all trials reviewed.

Methodological Issues

Study Design

All of the included trials were described as “randomized.” Four were in fact only pseudorandomized (i.e, treatment was allocated based alternate medical record numbers or birth dates, rather than by randomly generated numbers), which introduces the possibility of bias (Cardozo, Fysh, and Pearce, 1986; Hedén, Ingemarsson, Ahlström, et al., 1991; Katz, Yemini, Lancet, et al., 1983; Ohel, Rahav, Rothbart, et al., 1996). Two studies did not describe the method of randomization used (Egarter, Kofler, Fitz, et al., 1989; Herabutya, Prasertsawat, Tongyai, et al., 1992).

As discussed above and pointed out by Crowley (2000), the practical and ethical difficulties of blinding clinicians to either the target intervention or the results of antepartum testing results in an inherent bias against expectant management. Abnormal antenatal monitoring could influence a clinician’s thresholds for performing a cesarean section, either by making the diagnosis of “fetal distress” more likely or by a decreased willingness to augment labor aggressively.

In any trial of planned induction versus expectant management with antepartum testing, a certain proportion of women randomized to planned induction will go into spontaneous labor, while a proportion of women randomized to expectant management will have abnormal antepartum testing results; or, as observed in the Canadian Multicenter Post-term Pregnancy Trial (Hannah, Hannah, Hellmann, et al., 1992), patients or providers may request induction. These subjects are quite correctly analyzed in the groups to which they are randomized, rather than in accordance with the “treatment” received, since the trial is not comparing spontaneous delivery to induction, but instead, management strategies undertaken with the knowledge that some women will deliver spontaneously prior to scheduled induction, and some women will require (or request) induction during expectant management.

Outcome Measurement

All studies reported results for “hard” outcomes such as perinatal mortality and cesarean section rates. Reporting of other outcomes of interest was more variable. Many outcomes are subject to inherent difficulties with reproducibility and bias (e.g., the diagnosis of “fetal distress”), variability in operator preferences and skills (e.g., operative vaginal delivery rates), or are of uncertain long-term clinical significance (e.g., meconium-stained amniotic fluid in the absence of meconium aspiration, or Apgar scores). Other measures, such as patient preferences for different management strategies, longer-term neonatal outcomes, and vaginal and perineal trauma, would be of significant interest to patients, clinicians, and policymakers. We identified one cohort study published in 1991 which showed that patients’ preferences for induction versus expectant management changed with advancing gestation: 45 percent of women preferred conservative management at 37 weeks, compared with 31 percent at 41 weeks (Roberts and Young, 1991). Measurement of these preferences in light of data published subsequent to this study, and using methods developed and refined in the past decade, is needed. Detailed measurement of both medical and nonmedical costs is also lacking in the studies reviewed.

Comparability and Generalizability

The gestational age at which interventions were begun, as well as the methods used for induction and monitoring, varied between studies. Because variability in these methods may result in quite different outcomes, caution should be used when comparing outcomes that could possibly be affected by different methods of labor induction (such as cesarean section rates or time spent in labor) or different protocols for fetal monitoring (such as perinatal mortality) between studies. In addition, clinical management decisions may vary between practitioners. Especially in smaller trials, unequal distribution of different practitioners with different preferences and thresholds for management of labor may have resulted in some differences in outcomes.

Readers also must consider the degree to which these studies are generalizable to particular settings. If these methods or protocols are substantially different from those used in a particular setting, then the results may not be applicable. For example, the Canadian Multicenter Post-term Pregnancy Trial did not use prostaglandins for induction of women with abnormal antepartum testing (Crowley, 2000; Hannah, Hannah, Hellmann, et al., 1992). Use of prostaglandins could have changed the results by yielding lower cesarean rates in the induction arm through more successful inductions, as pointed out by Crowley (2000). On the other hand, the use of these agents in women with potentially compromised fetuses could have resulted in even higher cesarean section rates because of fetal compromise. A reanalysis of the Canadian trial using published success rates for prostaglandins found that more liberal use of these agents would still lead to a significantly higher cesarean section rate in the expectant management group because the cesarean section rate in the group induced because of abnormal testing would be substantially higher (Hannah, Huh, Hewson, et al., 1996).

Statistical Issues

Only the Canadian trial (Hannah, Hannah, Hellmann, et al., 1992) was sufficiently powered to detect differences in rare perinatal outcomes. Many of the remaining studies were also under-powered to detect differences in dichotomous outcomes.

Inappropriate summary measures and statistical tests were frequently used (e.g., mean parity or Bishop score, with comparison by t-test, when nonparametric statistics would be more appropriate). Variables that are frequently not normally distributed, such as length of stay and costs, also were not uniformly reported using medians, and the effect of a few outliers on comparisons was not evaluated.

Summary

Despite the methodological issues raised above, there is a consistent finding that perinatal mortality rates are lower with planned induction at 41 weeks or later compared with expectant management, a finding confirmed by a formal Cochrane meta-analysis (Crowley, 2000). Based on the observed absolute risk difference, the Cochrane meta-analysis estimated that 500 inductions were necessary to prevent one perinatal death.

It is interesting to consider these findings in light of our review of antepartum tests under Question 1. We found that there was a consistent qualitative pattern for the majority of tests studied, no matter what surrogate outcome for fetal compromise was used: sensitivity was lower than specificity, while negative predictive value was higher than positive predictive value. This implies that predictive values are driven by the relatively low rates of adverse outcomes associated with fetal compromise in prolonged pregnancy. If the measures used are valid surrogates for fetal compromise leading to stillbirth, then this should hold true for stillbirth as well: the negative predictive value of antepartum tests for stillbirth should be much greater than the positive predictive value. However, as the risk of stillbirth increases with increasing gestational age after 37 weeks, the negative predictive value should decrease, and the number of stillbirths in the setting of normal test results should increase.

Elective induction of labor results in a lower risk of stillbirth only after 41 weeks. One explanation for this, consistent with the findings on antepartum tests, is that the baseline risk of stillbirth is low enough prior to 41 weeks that the negative predictive value of antepartum tests is quite good. After 41 weeks, the increasing stillbirth risk results in poorer negative predictive value, so that one would expect excess stillbirths compared with elective induction.

Other perinatal outcomes did not appear to differ significantly between induction and expectant management groups.

Maternal outcomes did not differ between women managed with antepartum monitoring or with planned induction with the agents used in these studies. Specifically, overall cesarean section rates did not differ, either globally or in the subgroups analyzed by the Cochrane group (Crowley, 2000). If anything, cesarean section rates were lower in the induced groups.

Only one large trial reported costs, and based on 1992 costs and care provided, planned induction at 41 weeks was less expensive than expectant management with antepartum testing. However, because of significant changes in the technologies used and the economics of medicine in the interim, additional research is needed to better understand the cost implications of these two strategies. For example, if elective induction at 41 weeks is deemed to be preferable from a clinical standpoint for most patients, then a thorough analysis of the resources needed to institute such a policy would have to incorporate factors such as staffing on labor and delivery suites and postpartum units, since temporal patterns of patient flow may change.

Elective induction of labor at 41 weeks consistently appears to reduce the risk of stillbirth compared with management with antepartum testing, with no increase in maternal or neonatal risks, including no increase in cesarean section rates. At least 500 inductions would be needed to prevent one stillbirth. The societal tradeoffs in terms of economic resources used are unclear because of a lack of strong data applicable to current practice. Individual patients may have different values for these outcomes or perhaps for the “process” of childbirth—some women may place a very high value on avoiding any medical intervention.

Question 3: What are the benefits, risks, and costs of currently available interventions for induction of labor?

Approach

The evidence reviewed so far in this report suggests:

- The risk of perinatal death increases with advancing gestational age.
- There is no direct evidence that antepartum surveillance in prolonged gestation reduces perinatal morbidity or mortality. When surrogate measures are used as outcomes, the

consistent pattern of test characteristics for tests used in antepartum surveillance is for poor sensitivity but high negative predictive value, suggesting that false negative test results will become more likely as the underlying risk of adverse outcomes increases with advancing gestational age.

- Randomized trials show a reduction in perinatal mortality in women induced at 41 weeks gestation compared with women followed with antepartum testing, a finding consistent with increasing risk with advancing gestational age and with the observed patterns of test characteristics. Cesarean section rates are not increased in the elective induction arms of these studies.

Given that induction at 41 weeks appears to be effective in reducing mortality, data about the safest and most effective method of induction are needed in order to determine the optimal management strategy.

This section considers interventions designed to induce labor, including prostaglandin E_2 (PGE_2, or dinoprostone) gel (Prepidil®), PGE_2 tablets, PGE_2 insert (Cervidil®), misoprostol tablets, misoprostol gel, oxytocin, mifepristone, membrane sweeping, nipple stimulation, and other treatments. These methods are used either as primary methods of induction or as adjunctive methods in oxytocin induction. We limited our review to studies where the induction method was randomly assigned and compared with either placebo or a different induction method, and where at least some of the subjects were induced for an indication related to prolonged pregnancy. In this section, we also consider active interventions performed in the ambulatory setting at or near term that are designed to reduce the proportion of women reaching “postdates” or “postterm.”

In addition to the results of our review, we report summary conclusions based on meta-analyses performed for the Royal College of Obstetricians and Gynaecologists’ (RCOG) recent guideline on induction of labor (Royal College of Obstetricians and Gynaecologists, 2001) in collaboration with the Cochrane Collaboration.

Results

Castor Oil

We identified one randomized trial of castor oil used at term to promote spontaneous labor. Garry, Figueroa, Guillaume, et al. (2000) randomized women to 60 mg castor oil given orally in apple or orange juice (n = 52) or no treatment (n = 48). Mean gestational age was 284.4 ± 4.2 days in the castor oil group and 284.7 ± 3.6 days in the no treatment group. In the castor oil group, 57.7 percent of the subjects were in labor within 24 hours compared with 4.2 percent in the no treatment group ($p < 0.001$). Cesarean section rates were 19.2 percent in the castor oil group and 8.3 percent in the no treatment group ($p = 0.20$), but the study was underpowered to detect this difference or differences in rare outcomes such as uterine rupture. Of note, all women in the castor oil group experienced nausea. Other outcomes, such as proportion of women induced for other reasons or neonatal outcomes, were not reported.

The RCOG guideline (Royal College of Obstetricians and Gynaecologists, 2001) did not address castor oil. The most recent Cochrane review on the topic (Kelly, Kavanagh, and Thomas,

2001) identified the article cited above (Garry, Figueroa, Guillaume, et al., 2000) and reached conclusions similar to our own.

Breast Stimulation

We identified two studies that evaluated the use of breast stimulation in promoting the onset of labor near term and one that evaluated breast stimulation as a method of induction. Elliot and Flaherty (1984) randomized 100 women to either breast stimulation (manual stimulation of the nipple and areola for 15 minutes, alternating breasts, for a total of 1 hour at a time, three times daily) beginning at 39 weeks or a control pelvic examination; women in the control group were asked to abstain from sexual intercourse and avoid breast stimulation. Both groups were reevaluated at 42 weeks. Women with Bishop scores of 8 or greater were induced; others were followed with contraction stress tests. Five women in the breast stimulation group reached 42 weeks, compared with 17 in the control group; significance testing was not performed. Women in the breast stimulation group were significantly less likely to be induced after 42 weeks. The study was underpowered to detect differences in important outcomes, especially for the subgroup of women beyond 42 weeks.

Kadar, Tapp, and Wong (1990) randomized women at 39 weeks to either daily unilateral manual nipple stimulation “for as long as was practically feasible” (n = 60) or to no nipple stimulation (n = 76). There were no significant differences in any of the outcomes reported, including the proportion going into spontaneous labor, postterm deliveries, or median duration of pregnancy. Survival analysis showed that duration of pregnancy was related only to gestational age at enrollment and Bishop score. The authors also noted that adherence to the prescribed regimen was poor: 70 percent of the women assigned to the nipple stimulation group either failed to perform nipple stimulation at all or did so for less than 2 hours total during the entire study.

Chayen, et al., compared nipple stimulation using an electric breast pump to oxytocin as a method of induction (Chayen, Tejani, and Verma, 1986). In this study, only 29 percent of the inductions were for prolonged pregnancy. Thirty subjects were induced initially with a breast pump, while 32 received oxytocin. Time to achieve regular contractions and adequate labor as documented by intrauterine catheter were significantly less in the breast pump group. Cesarean section rates were also lower (26.7 percent vs. 43.7 percent in the oxytocin group), although this difference was not significant. Patients in the oxytocin group were more likely to have a higher Bishop score at baseline. Results were not reported separately by parity or for the subgroup of women induced for prolonged pregnancy.

In summary, because of lack of significance testing, poor compliance, or lack of power, the available randomized trials do not allow conclusions to be drawn about the effectiveness of breast stimulation in promoting labor or as a method of induction. The RCOG guideline (Royal College of Obstetricians and Gynaecologists, 2001) did not address this topic.

Relaxin

We identified three randomized trials of relaxin. Evans, Dougan, Moawad, et al. (1983) randomized women at 41 weeks gestation scheduled to undergo oxytocin induction of labor to intracervical or vaginal insertion of 4 mg relaxin (n = 10), 2 mg relaxin (n = 13), or placebo (n = 14); if the patient reached 42 weeks gestation, then labor was induced. No significant differences in any parameters, including days to admission, spontaneous labor, or time to delivery, were noted. There were trends towards a shorter time to delivery in the relaxin groups, but the study was underpowered to detect a difference for this outcome.

Bell, Permezel, MacLennan, et al. (1993) randomized women scheduled for induction for prolonged pregnancy to intravaginal 1.5 mg recombinant human relaxin (n = 18) or placebo (n = 22). No significant differences in any outcomes were reported. The authors noted that a low dose was deliberately chosen to help establish a safety profile for relaxin.

` Brennand, et al., randomized women between 37 and 42 weeks, "most" of whom were being induced for pregnancy-induced hypertension or prolonged pregnancy, to placebo or 1 mg, 2 mg, or 4 mg of recombinant relaxin (Brennand, Calder, Leitch, et al., 1997). There were no significant differences in any outcome except for slightly elevated baseline fetal heart rates after relaxin.

In summary, there are insufficient data available on relaxin to draw any conclusions about its safety or efficacy in induction of labor in women with prolonged pregnancy.

Sweeping of the Membranes

We identified 12 trials evaluating the efficacy of sweeping (or "stripping") of the membranes, 11 designed to evaluate the use of this intervention to promote spontaneous labor and reduce the need for induction and one in which it was used as a method of induction. In general, sweeping the membranes involves inserting a finger into the cervix and rotating the finger in the plane between the fetal membranes and the cervix and lower uterine segment. Details of the techniques used varied between studies and are described for each study in Evidence Table 3. Table 27 summarizes the 11 trials of membrane sweeping as a labor promoter.

All studies except one consistently showed higher rates of labor within a predefined time period, usually 1 week, in women randomized to active membrane sweeping. The proportion of women induced was also consistently lower in groups randomized to membrane sweeping. No differences in adverse outcomes, including infection or bleeding, were noted in any study. Level of patient discomfort during the procedure was not assessed in any study.

The one study that did not show a difference in outcomes (Crane, Bennett, Young, et al., 1997) was different from the other trials in several ways. Membrane stripping was performed only once. Patients in the stripping group were more likely to be nulliparous and to have lower Bishop scores. Stratified analyses and logistic regression did not show significant effects, but it is possible that the smaller sample size in these subgroups limited power. In addition, a survival analysis showed a decrease in the median time from enrollment to delivery (6.5 days for stripping, compared with 8 days for controls), but this difference was not significant.

In the one study in which membrane sweeping was used as an adjunct to induction of labor, Boulvain, et al., randomized women to sweeping of the membranes (n = 99) or vaginal examination only (n = 99) prior to induction of labor for "nonurgent" indications (Boulvain, Fraser, Marcoux, et al., 1998). Eighty-five percent of the patient population was induced for prolonged pregnancy. Mean time from randomization to onset of labor was significantly shorter in the sweeping group (76 hours vs. 98 hours; p = 0.01), but no significant differences were seen in other outcomes except patient discomfort (odds ratio [stripping vs. control], 2.52; 95 percent confidence interval [CI], 1.60 to 3.99), bleeding, and painful contractions without labor.

In summary, in all but one study, sweeping the membranes consistently promoted labor at term and reduced the incidence of induction for prolonged pregnancy. As with the majority of the interventions reviewed in this report, there are no data on patient preferences for this intervention. One study found that women who undergo membrane stripping are more likely to experience discomfort, bleeding, and painful contractions without labor compared with controls. Another issue is that the majority of studies excluded women whose cervices would not allow introduction of the examiner's finger; thus, the conclusions described are applicable only to those pregnant women at term whose cervices are dilated enough to allow introduction of an examiner's finger.

Similar findings have been reported in a Cochrane review (Boulvain and Irion, 2001) and incorporated into the RCOG guidelines (Royal College of Obstetricians and Gynaecologists, 2001).

Mechanical Devices

We identified two randomized trials of the use of mechanical devices such as Foley catheters, which are inserted into the cervix and then inflated. Atad, et al. (Atad, Hallak, Auslender, et al., 1996) compared 3 mg PGE_2 gel (n = 30), oxytocin (n = 30), and a double-balloon catheter invented by one of the investigators (n = 35). Patients in the first two groups crossed over to the catheter arm if the Bishop score was ≤ 4 at 12 hours, while patients in the catheter group received PGE_2 if the Bishop score was ≤ 4 at 12 hours. More patients in the catheter group had cervical dilation > 3 cm after 12 hours (86 percent vs. 23 percent in the oxytocin group and 50 percent in the PGE_2 group; $p < 0.01$). Both PGE_2 and the balloon device had higher rates of vaginal delivery (PGE_2, 70 percent; catheter, 77 percent; oxytocin, 27 percent) and lower rates of cesarean section among patients with cervical dilation after the initial intervention (PGE_2, 13 percent; catheter, 18 percent; oxytocin, 43 percent). Only 18 percent of the inductions in this study were for prolonged pregnancy.

Sciscione, et al., randomized 53 women to misoprostol and 58 to mechanical dilation with a 16 F Foley catheter with a 30 cc balloon (Sciscione, Nguyen, Manley, et al., 2001). There were no significant differences in change in Bishop score, vaginal delivery rates, or time to delivery in the two groups. Uterine tachysystole and passage of meconium were significantly more frequent in the misoprostol group. There was a trend towards higher cesarean section rates for nonreassuring fetal heart rate tracing in the misoprostol group (24 percent vs. 12 percent; $p = 0.09$), in a study where the sample size was determined based on change in Bishop score. Only 16 of 111 women in this study were induced for an indication of prolonged pregnancy.

In these two trials, mechanical devices appear to be comparable to prostaglandins in terms of delivery success, with lower rates of fetal heart rate tracing changes associated with frequent uterine contractions. As with membrane sweeping, applicability is limited to women whose cervix is dilated enough to allow introduction of a catheter. As with the majority of the other interventions reviewed, these studies also included relatively few women in the population of interest (prolonged pregnancy with no other risk factors) and were underpowered to detect differences in many important outcomes.

Mechanical devices alone are not addressed specifically in published Cochrane reviews or in the RCOG guideline (Royal College of Obstetricians and Gynaecologists, 2001).

Oyxtocin Dosing

We identified one randomized trial comparing two dosing regimens of oxytocin. Satin, Hankins, and Yeomans (1991) randomized women being induced for prolonged pregnancy to a “slow-dose” regimen (an initial dose of 2 mU/min, with increments of 1 mU/min at 30-minute intervals) or a “fast-dose” regimen (an initial dose of 2 mU minute with increases of 2 mU/min at 15-minute intervals). Induction failure was more likely in the slow-dose group (31 percent vs. 8 percent; $p < 0.05$). Time to delivery was shorter in the fast-dose group in both nulliparous women (9 hours vs. 15 hours; $p < 0.05$) and multiparous women (8 hours vs. 11 hours; $p < 0.05$). No significant differences were observed in other outcomes. There was a trend towards more hyperstimulation episodes requiring cessation of oxytocin in the fast-dose group, but the study was underpowered to detect a difference.

There is no formal comparison of oxytocin dosing regimens in published Cochrane reviews. The RCOG guideline development group reviewed dosing regimens in 11 trials of oxytocin with and without amniotomy. Their qualitative conclusions were: (1) lower dose regimens were not associated with an increase in operative delivery rates; (2) regimens with incremental rises in dose more frequently than every 30 minutes were associated with an increase in uterine hypercontractility; (3) lower dose regimens were not associated with an increase in specified delivery intervals; and 4) higher dose regimens were associated with an increase in the incidence of precipitous labor (Royal College of Obstetricians and Gynaecologists, 2001).

Prostaglandins

Of the randomized trials identified, 20 evaluated PGE_2 (dinoprostone) gel, five evaluated PGE_2 tablets, one evaluated the Cervidil® insert, one evaluated low-dose (2 mg) PGE_2 vaginal suppositories, and 22 examined misoprostol. Placement of the prostaglandin was either intravaginal (usually in the posterior fornix) or intracervical. The site of application is described for each study in Evidence Table 3 and in the text below.

PGE_2 gel in an ambulatory setting to reduce the need for induction. Five studies examined the effect of PGE_2 gel versus placebo (Buttino and Garite, 1990; Doany and McCarty, 1997; Lien, Morgan, Garite, et al., 1998; O'Brien, Mercer, Cleary, et al., 1995; Sawai, Williams, O'Brien, et al., 1991). Doany and McCarty (1997) randomized patients to one of four arms: (1) no membrane stripping and placebo gel; (2) no membrane stripping and PGE_2 gel; (3) membrane stripping and placebo gel; or (4) membrane stripping and PGE_2 gel. Gel was placed in the posterior vaginal fornix. PGE_2 gel without membrane stripping was not significantly different from placebo without stripping for any outcome. All patients in this study were 41 weeks or greater in gestational age.

Lien, et al., a randomized trial of intracervical PGE_2 gel ($n = 43$) versus placebo ($n = 47$) begun after 40 weeks, found no significant differences between the two arms in the interval from admission to delivery, cesarean sections, or maximum oxytocin dosage (Lien, Morgan, Garite, et al., 1998). For patients who presented with a Bishop score between 3 and 6, those who were randomized to PGE_2 gel were less likely to be induced than those treated with placebo gel.

Sawai, Williams, O’Brien, et al. (1991) randomized women at 41 weeks to either weekly PGE_2 gel in the posterior fornix ($n = 24$) or weekly placebo gel. Induction occurred if the Bishop score was greater than 9, in the event of abnormal fetal heart rate testing, or at 44 weeks. There were no significant differences in neonatal outcomes, cesarean section rates, length of labor, or time from randomization to admission between the two groups, but the study was underpowered to identify differences in most categorical variables.

Buttino and Garite (1990) randomized women at 41-6/7 weeks to either intracervical PGE_2 (n = 23) or placebo (n = 20). There were no significant differences in any outcome, including neonatal outcomes, cesarean section rate, or time to delivery. Cesarean section rates were lower in the PGE_2 group (21.7 percent vs. 35.0 percent), but the study was underpowered to detect a difference. Gestational age at delivery and time from randomization to delivery were not significantly different in the two induction groups.

O'Brien, et al., randomized women at 38-39 weeks to intravaginal PGE_2 gel (n = 50) or placebo (n = 50) daily for 5 days (O'Brien, Mercer, Cleary, et al., 1995). PGE_2 gel resulted in significantly fewer pregnancies going beyond 40 weeks (40 percent vs. 66 percent; $p < 0.016$), although not in the proportion of pregnancies reaching 42 weeks (4 percent vs. 6 percent). Induction rates were lower in the PGE_2 group (12 percent vs. 28 percent; $p = 0.08$).

PGE_2 gel as an adjunct to oxytocin. A randomized trial conducted by the National Institute of Child Health and Human Development (NICHD) Network of Maternal-Fetal Medicine Units (1994) compared induction between 41 and 42 weeks and expectant management. The induction group in this trial was split into two arms: intracervical PGE_2 gel plus oxytocin (n = 174) and placebo gel plus oxytocin (n = 174). No significant differences in neonatal or maternal outcomes, including cesarean section rates, were detected between the two groups. Sample size estimates for this trial were based on perinatal morbidity and mortality and maternal mortality.

Rayburn, et al., compared intracervical PGE_2 gel (n = 55) to placebo (n = 63) prior to induction of labor with oxytocin at 42 weeks (Rayburn, Gosen, Ramadei, et al., 1988). Overall cesarean section rates (18 percent with PGE_2 gel vs. 33 percent with placebo; $p < 0.05$) and mean time to delivery (5.5 hours vs. 9.5 hours with placebo; $p < 0.01$) were significantly lower with PGE_2 gel.

Chatterjee, et al., compared 2 mg PGE_2 gel to placebo (Chatterjee, Ramchandran, Ferlita, et al., 1991). Bishop scores were significantly improved in patients receiving the active gel; the study was underpowered to detect any other differences.

PGE_2 gel dosing. Voss, Cumminsky, Cook et al. (1996) compared the use of intracervical PGE_2 gel in three different dosing regimens: 0.125 mg (n = 79), 0.25 mg (n = 70), and 0.5 mg (n = 80). For each of the outcomes described (fetal heart rate abnormality, cesarean sections, mean change in Bishop score, hyperstimulation, and time to active phase labor/complete dilation/delivery), there was no significant difference noted for the various doses of PGE_2 gel. Only 31 percent of subjects in this study were induced for prolonged pregnancy.

MacKenzie and Burns (1997) compared a single vaginal dose of 2 mg PGE_2 gel, with amniotomy and oxytocin if no labor occurred within 14-20 hours of treatment, with 2 mg of PGE_2, followed by a second application in 6 hours if no labor occurred or if the Bishop score was less than 9. Sixty-eight percent of the patients in this trial were induced for prolonged pregnancy. The only significant difference noted was a shorter time to delivery in the two-dose group among multiparous women (mean 785 minutes vs. 927 minutes in the single-dose group).

Graves, et al., compared PGE_2 gel in doses of 1 mg, 2 mg, and 3 mg to placebo prior to induction with oxytocin (Graves, Baskett, Gray, et al., 1985). Eighteen percent of the inductions were for prolonged pregnancy. There was a significant increase in Bishop score after the active gel compared with placebo, but this effect was not dose-related. There was a dose-related increase in the proportion of women entering spontaneous labor after insertion of the gel. There was a trend toward more uterine hypercontractility with higher doses of the gel, although the study was underpowered to detect a significant difference. Other outcomes were not significantly different between the active and placebo groups, although the study lacked power to detect many differences.

PGE_2 gel versus PGE_2 tablets. One study compared 3 mg PGE_2 tablets to 2 mg PGE_2 gel (Mahmood, 1989). The gel formulation required fewer applications and resulted in greater changes in Bishop score and shorter time to onset of labor than did tablets.

PGE_2 gel versus oxytocin. Two studies were identified that compared the administration of PGE_2 gel to induction by oxytocin infusion. In the first study (Papageorgiou, Tsionou, Minaretzis, et al., 1992), cesarean section for cephalopelvic disproportion and fetal distress, vacuum suction, and hyperstimulation were not statistically different in women randomized to intracervical PGE_2 (n = 83) or oxytocin (n = 82) for induction of labor after 41 weeks. Two outcomes did show benefit to the use of PGE_2 gel. First, babies were less likely to have an Apgar score < 7 at 5 minutes when the cervices of the mother were ripened by PGE_2 gel as opposed to those induced with oxytocin. Also, patients were more likely to be delivered vaginally if ripened by PGE_2 gel (89 percent vs. 71 percent). All subjects in this study had a gestational age of at least 41 weeks.

The second study (Misra and Vavre, 1994) compared administration of intracervical PGE_2 gel (n = 80) with oxytocin (n = 72). Rates of cesarean deliveries were decreased with PGE_2 in primigravidas only (26.3 percent with PGE_2 vs. 47.2 percent with oxytocin; $p < 0.01$). Women in this study were induced for a variety of indications, with a mean gestational age less than 40 weeks.

Placement of PGE_2 gel. One study examined the effect of placement of PGE_2 gel in the posterior vaginal fornix versus in the endocervical canal (Kemp, Winkler, and Rath, 2000). The outcomes that showed significance indicated that patients who received gel administered in the posterior vaginal fornix were more likely to deliver earlier (15.7 hours vs. 19.1 hours) and more likely to deliver in 24 hours (81.6 percent vs. 67.8 percent). In this study, 32.9 percent of the posterior fornix group were induced for prolonged pregnancy (more than 10 days past the estimated date of confinement), and 29.2 percent of the intracervical group were 10 days beyond term.

PGE_2 gel versus membrane stripping. Two studies compared outcomes between PGE_2 gel administration and membrane stripping. In Magann, et al., three groups were randomly assigned to treatment at 41 weeks (Magann, Chauhan, Nevils, et al., 1998). One group received daily intracervical administration of PGE_2 gel, another received daily membrane stripping, and the third group received a daily "gentle cervical examination." Patients in all three groups were induced if the Bishop score became ≥ 8, or at 42 weeks. Inductions at 42 weeks were significantly lower in the two active treatment groups (17 percent in the sweeping group and 20 percent in the PGE_2 group, compared with 60 percent in the controls). Cesarean section rates were higher in the PGE_2 group (8/35, or 23 percent, vs. 5/35, or 14 percent, in the other two groups), a relative risk of 1.6 (95 percent CI, 0.58 to 4.41).

In Doany and McCarty (1997), the effects of membrane stripping, PGE_2 gel (placed in the posterior vaginal fornix), and a combination of the two therapies were evaluated. Patients were randomized at 41 weeks to one of 4 groups: (1) membrane stripping and placebo gel;

(2) membrane stripping and PGE_2 gel; (3) "control" cervical exams and placebo gel; or (4) "control" exams and PGE_2 gel. Gestational age at delivery was significantly lower in the group with both active treatments (median, 290 days vs. 294 days in the two groups with one placebo and 297 days in the group with two placebos; $p = 0.005$). There was a trend towards a higher cesarean rate in the group with both active treatments (11 percent versus 8 percent in the two single-agent arms and 4 percent in the double-placebo group; $p = 0.08$).

These two studies suggest that PGE_2 is equivalent to membrane stripping in terms of promoting labor. In both studies, PGE_2 was associated with higher cesarean section rates, although these differences were not statistically significant. Larger studies would be needed to detect a difference in cesarean rates.

PGE_2 inserts. Only one study was identified that examined the efficacy of the Cervidil® vaginal insert (Wing, Ortiz-Omphroy, and Paul, 1997). This trial compared the Cervidil® insert (10 mg in a timed-release preparation) to 25 µg of misoprostol administered every 4 hours to a maximum of six doses. There were no significant differences between the two groups in neonatal or maternal outcomes. While the mean time to delivery was the same between the two groups, the misoprostol dosing every 4 hours showed a lower rate of tachysystole than the Cervidil® insert.

PGE_2 suppositories. One study evaluated the use of 2 mg intravaginal PGE_2 suppositories (n = 38) versus placebo suppositories (n = 42) self-administered by the patient on an outpatient basis beginning at 41 weeks (Sawai, O'Brien, Mastrogiannis, et al., 1994). The patients in the PGE_2 arm used fewer suppositories and were admitted for delivery at earlier gestational ages. This resulted in lower antepartum testing charges (mean $477 vs. $647 with placebo; $p = 0.001$). There was a trend towards lower cesarean section rates in the PGE_2 group (2.6 percent vs. 14.3 percent in the placebo group), although this difference was not significant.

In summary, vaginal or intracervical PGE_2 was consistently more effective in achieving cervical ripening or delivery within a specified time period compared with placebo or oxytocin. Cesarean section rates were lower or similar in women treated with PGE_2. There were no differences in perinatal or maternal morbidity or mortality.

Similar findings were reported in the review conducted for the RCOG guideline group. Based on their "conflated" analysis of trials comparing PGE_2 with oxytocin with or without amniotomy, the guidelines recommended PGE_2 as the treatment of choice for induction in women with intact membranes (Royal College of Obstetricians and Gynaecologists, 2001).

Misoprostol

Misoprostol tablets versus placebo. Only one study was identified that compared misoprostol with placebo prior to scheduled induction (Fletcher, Mitchell, Simeon, et al., 1993). A dose of 100 µg misoprostol (n = 32) was found to be more effective than placebo (n = 31). Time from induction to delivery was lower with misoprostol (22 hours vs. 32 hours), as was cesarean section rate (3 percent vs. 10 percent), although these differences were not statistically significant. The mean Bishop score was increased for patients treated with misoprostol. Only one-third of the randomized patients were induced for prolonged pregnancy.

Misoprostol tablets versus PGE_2 gel. Table 28 summarizes results from the 10 studies that compared intravaginal misoprostol tablets with intracervical or intravaginal PGE_2 gel (Buser, Mora, and Arias, 1997; Chuck and Huffaker, 1995; Fletcher, Mitchell, Frederick, et al., 1994; Gottschall, Borgida, Mihalek, et al., 1997; Herabutya, Prasertsawat, and Pokpirom, 1997; Howarth, Funk, Steytler, et al., 1996; Kadanali, Küçüközkan, Zor, et al., 1996; Mundle and Young, 1996; Varaklis, Gumina, and Stubblefield, 1995; Wing, Jones, Rahall, et al., 1995).

The studies examined a range of doses and frequency of dosing with similar results. The time from induction to delivery was consistently shorter in patients treated with misoprostol, both for all patients and for those with vaginal delivery. With one exception, misoprostol was shown to cause higher frequency of uterine hyperstimulation, hypertonus, or tachysystole, although studies were often underpowered to detect significant differences in these outcomes. All studies indicated that misoprostol was an effective agent for cervical ripening and induction, often more effective than PGE_2 gel, and showed no significant difference in the rates of cesarean section. One study (Buser, Mora, and Arias, 1997) showed an increase in cesarean section rates for patients treated with misoprostol; this was attributable to significantly higher rates of nonreassuring fetal heart rate patterns. Of note, the majority of subjects in these studies were not women being induced for prolonged pregnancy.

Misoprostol dosing studies. Two studies evaluated various dosing regimens for misoprostol. In Farah, et al., intravaginal administration of doses of 25 µg versus 50 µg every 3 hours was evaluated (Farah, Sanchez-Ramos, Rosa, et al., 1997). In this study, the incidences of hyperstimulation, tachysystole, and cord pH < 7.16 were greater in patients on the 50-µg regimen. In comparison, patients given 50 µg every 3 hours were more likely to have shorter start-to-delivery times and more vaginal deliveries.

In Wing and Paul (1996), the dosing regimen was 25 µg given either every 3 or 6 hours. Patients randomized to the 6-hour regimen had longer times to delivery, more frequently required oxytocin augmentation, and had more failed inductions than those on the 3-hour regimen.

Misoprostol versus oxytocin. Three studies compared the effect of intravenous oxytocin with intravaginal misoprostol (Escudero and Contreras, 1997; Kramer, Gilson, Morrison, et al., 1997; Sanchez-Ramos, Kaunitz, Del Valle, et al., 1993). Although the studies used varying dosages of misoprostol, the conclusions were similar. Patients treated with misoprostol had shorter induction-to-delivery times, more vaginal deliveries, and fewer cesarean deliveries for dystocia. Most studies also indicated that higher rates of uterine tachysystole were associated with misoprostol, and studies with higher doses of misoprostol had higher rates of tachysystole. Kramer, et al., found that patients treated with misoprostol also were less likely to use epidural anesthesia, and the costs associated with misoprostol induction were less than for patients induced by oxytocin (Kramer, Gilson, Morrison, et al., 1997). In this study, the costs associated with misoprostol treatment often excluded the cost of epidural anesthesia, longer length of stay (associated with induction), and fewer cesarean deliveries.

Method of delivery with misoprostol. Two studies examined the effect of various methods of delivery for the dosing of misoprostol. Srisomboon, et al., evaluated the effect of 100 µg of misoprostol given intracervically versus intravaginally (after dissolution of the misoprostol pill into an inert gel) (Srisomboon, Piyamongkol, and Aiewsakul, 1997). There were no significant differences found between the two methods of administration in terms of change in Bishop score, interval from administration to delivery, route of delivery, or perinatal outcome. Rates of uterine tachysystole were similar in the two groups. This study noted that spillage of gel out of the cervix was observed in 70 percent of patients receiving intracervical misoprostol. The investigators concluded that the rates of efficacy between the two methods were similar, and that intravaginal administration was more convenient. Thirty-four percent of the inductions in this study were for prolonged gestation.

Toppozada, Anwar, Hassan, et al. (1997) evaluated the effects of oral versus vaginal misoprostol. Forty patients were randomized to 100 µg every 3 hours administered via the oral or vaginal route. Patients were more likely to be induced successfully via the vaginal route in a shorter interval at a lower dose but were also more likely to experience abnormal fetal heart rate patterns and higher rates of uterine hyperstimulation. The proportion of subjects induced for prolonged pregnancy was not reported in this study.

Misoprostol tablet versus PGE_2 tablet. Four studies were identified that evaluated the effects of intravaginal PGE_2 tablets to intravaginal misoprostol tablets (Chang and Chang, 1997; Fletcher, Mitchell, Frederick, et al., 1994; Lee, 1997; Surbek, Boesiger, Hoesli, et al., 1997). While the dosing regimens for the studies differed, the conclusions were similar. Patients treated with misoprostol were found to have shorter intervals between insertion and delivery, had higher mean Bishop scores 12 hours after administration, and were more likely to deliver in 24 hours. Three of the four studies concluded that misoprostol was a more effective and efficient drug for induction than PGE_2. No significant differences in perinatal outcomes were noted.

Misoprostol versus PGE_2 insert (Cervidil®). One study compared the effects of the Cervidil® vaginal insert with misoprostol (Wing, Ortiz-Omphroy, and Paul, 1997). Patients randomized to treatment with Cervidil® had higher rates of tachysystole and abnormal fetal heart rate patterns. There were no significant differences in perinatal outcomes. Patients treated with misoprostol had shorter intervals from start to delivery than those treated with Cervidil®, but this difference was not significant. This study concluded that misoprostol was as effective as Cervidil®, but that the incidence of uterine tachysystole was significantly lower with misoprostol.

In summary, the majority of the randomized trials of misoprostol showed that misoprostol was more effective in achieving vaginal delivery within 24 hours than were other induction agents. However, misoprostol was also more likely to result in uterine hypercontractility, a not unsurprising correlate of efficacy. All the studies reviewed were underpowered to detect clinically relevant differences in many important outcomes, particularly those having to do with safety. Similar conclusions have been reached by recent Cochrane reviews on misoprostol (Alfirevic, Howarth, and Gaussmann, 2000; Hofmeyr and Gulmezoglu, 2001).

Mifepristone

We identified five studies that compared the efficacy of the progesterone receptor antagonist mifepristone (RU-486) to placebo. Unlike many of the studies discussed above, three of the five focused on patients primarily induced for prolonged pregnancy. All five studies indicated that mifepristone was effective in ripening the cervix. Wing, et al., using 200 mg mifepristone, found significantly more deliveries and vaginal deliveries within 48 hours and a shorter time to delivery with mifepristone compared with placebo; subgroup analysis showed that these effects were primarily due to the effect in nulliparas (Wing, Fassett, and Mishell, 2000). There were trends towards more abnormal fetal heart rate tracings in labor and more infants with Apgar scores less than 7 at 1 and 5 minutes in the mifepristone group, but these trends did not reach statistical significance.

Three studies evaluated patients who were treated with 400 mg mifepristone versus placebo. In Stenlund, Ekman, Aedo, et al. (1999), the time to onset of labor was shorter and the proportion of patients in labor within 48 hours was significantly greater (81.8 percent vs. 27.3 percent) in the mifepristone group. Median Apgar scores at 1 minute were lower in the mifepristone group, but there were no differences in Apgar scores at 5 or 10 minutes. With only 36 subjects, this study was underpowered to detect differences in many outcomes.

In Giacalone, et al., time to onset of labor and time to vaginal delivery were significantly shorter in the mifepristone group (Giacalone, Targosz, Laffargue, et al., 1998). There were trends towards lower Apgar scores at 1 minute and lower cord pH values, but these were nonsignificant; again, the study was severely underpowered to detect differences in many important clinical outcomes, including cesarean section rate.

In Frydman, et al., the proportion of women going into spontaneous labor, the proportion with Bishop scores less than 4 at presentation for induction, and the mean randomization-to-delivery time were all significantly less in the mifepristone group (Frydman, Lelaidier, Baton-Saint-Mleux, et al., 1992). There were no significant differences in other outcomes and no other trends. Again, the study was underpowered to detect differences in safety-related outcomes. Forty-eight percent of the patients were induced for “postdate” pregnancy.

Elliott, et al., performed a dose-response study comparing placebo with 50 mg and 200 mg of mifepristone in nulliparous women, the “majority” of whom were being induced for prolonged pregnancy (Elliott, Brennand, and Calder, 1998). When a combined outcome measure of either spontaneous labor within 4 days or Bishop score of ≥ 6 at induction was used as the measure of efficacy, there were significant improvements with mifepristone in a dose-related manner. However, mifepristone was also associated in a dose-related manner with significantly more cases of fetal distress in labor and neonatal jaundice. In addition, cesarean rates were significantly lower with 50 mg of mifepristone than with placebo but higher with 200 mg than with placebo ($p = 0.07$), a difference that appears to be attributable to a higher incidence of cesarean delivery for fetal distress in the 200-mg group.

In summary, mifepristone appears to be superior to placebo in terms of achieving labor or cervical ripening within a specified time, but there are consistent trends towards fetal compromise during labor in women who receive mifepristone. Inadequate power to detect potentially important differences in safety argue against the use of mifepristone for induction of labor in prolonged pregnancy outside of research protocols at the present time.

A Cochrane review on this topic found similar evidence of efficacy (Neilson, 2001). Neonatal outcomes were not reported in enough studies to allow conclusions about safety.

Methodological Issues

In reviewing the literature on induction agents, numerous methodological problems consistently reduced our ability to draw conclusions about the benefits and risks of these agents in managing women with prolonged pregnancy. Some of these problems concerned study design; others related to statistical issues.

The following observations may be made about study design:

- Patient population: The majority of the studies evaluating the efficacy of different interventions for induction of labor included subjects with a range of indications for induction and did not report results separately for those women induced because of prolonged pregnancy. This has several implications. First, it is possible that the responsiveness of the uterus and cervix (even with comparable Bishop scores) to a given agent might be quite different between a woman at 37 weeks with preeclampsia and a woman at 42 weeks with no medical complications, leading to different estimates of efficacy. Second, risks for fetal compromise might also be quite different between a woman at 37 weeks with preeclampsia compared with a woman at 41 weeks with no medical complications compared with a woman at 42 weeks with oligohydramnios. The two groups of interest in this report are women induced solely because of prolonged gestation and women induced because of abnormal antepartum surveillance in prolonged gestation. The majority of the literature does not allow us to draw conclusions about the risks and benefits of particular induction agents in these two groups. Several studies also noted differences in outcomes between nulliparous and parous women; the majority failed to stratify results by parity.

- Choice of primary outcomes: Of those studies that stated an a priori sample size estimation, most based it on time-related outcomes, such as time to delivery, time to vaginal delivery, or proportion of subjects delivering within 24 or 48 hours. Although these certainly are important outcomes, sample size estimates based on these types of outcomes will inevitably lead to studies that are underpowered to detect clinically relevant differences in other important outcomes, such as perinatal morbidity or cesarean section rates. This was found throughout the misoprostol literature, where there were consistent trends towards higher rates of uterine tachysystole, hyperstimulation, and nonreassuring fetal heart rate tracings, but most studies were underpowered to detect the differences. Studies that based their sample size estimates on changes in the Bishop score failed to account for the inherent intra- and interobserver variability of this measurement; accounting for this would have led to larger sample sizes.

- Variability in clinical management: As with most of the studies reviewed for this report, variability in clinical management of labor may have resulted in differences in many outcomes, especially cesarean section rates, which make comparisons across studies difficult.

- Patient preferences: Consistently, time to delivery was chosen as an important outcome variable. Not surprisingly, more rapid times to delivery were associated with intermediate markers of fetal compromise or potential fetal compromise. Time to delivery is an important resource use issue. However, given the potential tradeoffs, collection of patient-oriented outcomes (preferences for the tradeoff of time in labor vs. risk of fetal compromise, for example) would be a valuable adjunct to these studies.

- Cost data: Few studies reported cost data. Those that did frequently failed to account for all medical costs and focused only on pharmacy-related costs. This lack of data prevents estimation of cost-effectiveness.

The following observations are made about statistical issues:

- Sample size: As stated above, the choice of primary outcome variable often inhibited the ability of trials to detect potentially clinically relevant differences in important outcomes. This is particularly true for rare but clinically important outcomes such as uterine rupture. There are case reports of uterine rupture occurring in women without previous uterine surgery after induction with misoprostol (Bennett, 1997; Blanchette, Nayak, and Erasmus, 1999); whether the risk of this event is higher in women induced with misoprostol compared with other medications is unclear, since denominator data are not available. However, the lack of statistical power to detect categorical events in the majority of randomized trials of induction agents is a major limitation to interpretation of this literature.

- Choice of statistical tests: Inappropriate statistical tests (e.g., means for integer variables such as parity, Apgar or Bishop score, or for nonnormally distributed variables, such as length of stay or time in labor) were frequently used. Use of these summary measures could potentially lead to false conclusions about the comparability of groups at either baseline or after intervention.

Summary

Based on the above review, we conclude the following:

- The majority of randomized trials of induction agents where a priori sample size estimates were performed are powered based on detecting a difference in outcomes such as time to delivery. This results in a lack of power to detect clinically meaningful differences in categorical outcomes that are less common. This lack of power precludes drawing definite conclusions about the relative safety of different agents.

- Castor oil given at term appears to be effective in promoting labor, with a consistent side effect of maternal nausea; whether other outcomes of interest are affected is unclear.

- Manual nipple stimulation at term may promote labor; effectiveness may be dependent on the protocol used and patient ability to adhere to the protocol. Currently available data are insufficient to draw conclusions.

- Data on the effectiveness of electrical breast stimulation as a method for inducing labor in prolonged gestation are inconclusive because of small sample size and a low proportion of subjects induced for an indication of prolonged pregnancy.

- Data on the safety and effectiveness of relaxin are limited and no conclusions can be drawn.

- Sweeping of the membranes at or near term is effective in promoting labor and reducing the incidence of induction for prolonged gestation.

- In general, there is a tradeoff between the effectiveness of induction agents when effectiveness is defined in terms of achieving delivery and shortening the time to delivery on the one hand, and risks of uterine tachysystole, hyperstimulation, and potential fetal compromise on the other. In increasing order of effectiveness, slow-dose oxytocin is followed by fast-dose oxytocin; PGE_2 appears more effective than oxytocin, and misoprostol is more effective than PGE_2. The heterogeneity of the patient populations in the published literature prohibit definitive conclusions about the benefits and risks of these agents in the setting of induction of labor in prolonged pregnancy, either for women induced electively or for women with abnormal fetal surveillance.

- Mifepristone (RU-486) is consistently effective in reducing the time to labor and the time to delivery in women after 41 weeks. However, all three published trials reported nonsignificant trends towards higher rates of intermediate markers of fetal compromise, including abnormal fetal heart rate tracings and low Apgar scores.

- Data on costs are insufficient to allow conclusions about cost-effectiveness.

Question 4: Are the epidemiology and outcomes of prolonged pregnancy different for women in different ethnic groups, different socioeconomic groups, or in adolescent women?

Approach

We approached this question in two ways. First, in all the articles we reviewed, we searched for data on differences in either the epidemiology or outcomes of prolonged pregnancy in different ethnic groups, different socioeconomic groups, and different age groups. Second, we reviewed published data from birth certificates (Ventura, Martin, Curtin, et al., 2000) and from the 1997 Nationwide Inpatient Sample (NIS) (Nationwide Inpatient Sample [NIS], 1997). The NIS is part of the Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project (HCUP). HCUP collects discharge data from a stratified sample of approximately 20 percent of U.S. hospitals. Using ICD-9 codes, we divided all deliveries into "preterm" (644.2x), prolonged (645.x), and term (all other delivery codes). We examined differences in outcomes between coded ethnic groups (white, black, Hispanic, Asian/Pacific Islander, Native American, and "other") and by insurance status (Medicare, Medicaid, private/health maintenance organization [HMO], self-pay/no insurance, "no charge," and "other") within these categories.

Results

Racial and Ethnic Differences: Literature Review

We did not identify any articles that specifically addressed differences in the epidemiology or outcomes of prolonged pregnancy in different ethnic groups.

Racial and Ethnic Differences: Primary Data

Birth certificate data. Table 29 summarizes total births, with percentages of infants born after 40 weeks, 41 weeks, and 42 weeks, from 1998 birth certificate data reported to the National Center for Health Statistics (NCHS), by race of mother (Asian or Native American data are not available in the published report). The proportions reported were calculated from the absolute numbers provided in the NCHS report. Table 29 also illustrates the proportion of live births after 42 weeks that were low birthweight (less than 2,500 grams) or macrosomic (greater than 4,000 grams).

Taking into account the limitations of birth certificate data, there are some interesting findings:

- Live births between 40 and 42 weeks were less common for non-Hispanic black women than for non-Hispanic white women, which may be partly due to an increased risk of preterm birth among non-Hispanic blacks (17.5 percent vs. 10.2 percent in non-Hispanic whites). However, the proportion of births after 42 weeks is strikingly similar in all groups.

- The weight distribution among infants born after 42 weeks is also strikingly different between groups, with non-Hispanic black women having a two-fold increase in low birthweight infants and a substantially lower incidence of macrosomic infants.

Hospital discharge data. Table 30 shows the percentage distribution of selected discharge diagnoses in the subset of women with a primary discharge diagnosis of prolonged pregnancy, by coded ethnic group. Total raw discharges in the NIS with this diagnosis were 57,814, or 7.2 percent of the total pregnancy-related discharges. Again, black women were more likely than women in other ethnic groups to have a diagnosis of restricted fetal growth and were less likely to have a diagnosis of macrosomia than white or Hispanic women. Black women also were more likely to have diagnoses of fetal distress and oligohydramnios. Interestingly, they also were somewhat more likely to have a diagnosis of shoulder dystocia than white or Hispanic women. Asian/Pacific Islander women were more likely to have diagnoses of macrosomia but less likely to have perineal trauma of any kind. Potential explanations for this observation include a higher cesarean section rate in Asian/Pacific Islander women, differences in the pelvic floor, or dynamics of labor which make perineal trauma less likely.

Both the NIS data and birth certificate data suggest that black women are more likely to have low birthweight infants after 42 weeks than white or Hispanic women. Diagnoses such as oligohydramnios and fetal growth restriction are also more common in black women. All three of these diagnoses are consistent with declining uteroplacental function. There were a limited number of fetal deaths in the NIS data set, with racial data missing from over half.

Socioeconomic Groups: Literature Review

We did not identify any articles that specifically addressed differences in the epidemiology or outcomes of prolonged pregnancy in different socioeconomic groups.

Socioeconomic Groups: Primary Data

Table 31 shows the percentage distribution of coded discharge diagnoses by payer status of women with a diagnosis of prolonged pregnancy. Women with private or HMO insurance coverage were less likely than women with Medicaid or no insurance to have diagnoses of intrauterine growth restriction or oligohydramnios.

Age Differences: Literature Review

We did not identify any articles that specifically addressed differences in the epidemiology or outcomes of prolonged pregnancy in either adolescent women or women in their later reproductive years.

Methodological Issues

Data Quality Issues

The accuracy of the dating recorded on birth certificates is unconfirmable, at best. Therefore, it is unclear whether the observed trends in racial differences in the distribution of birthweight after 42 weeks, and the observed lack of difference in the proportion of all pregnancies that reach 42 weeks, are real or simply random error introduced by variable quality of dating.

Similarly, criteria for a diagnosis of prolonged pregnancy, as well as for many of the other diagnosis codes, may vary between hospitals. Data for racial and payer codes were missing for many of the coded complication diagnoses. If codes are not recorded systematically in some hospitals, this may result in misleading patterns.

Statistical Analysis

Because of concerns with data quality, we did not perform formal tests of significance or multivariate analyses. Given the consistent patterns for some observations seen in the two data sets, more detailed analysis of more complete data sets is warranted.

Summary

The current published literature on the epidemiology and management of prolonged pregnancy does not provide information on the potential effects of race and ethnicity, socioeconomic status, or age on the incidence and outcomes of prolonged pregnancy. Given that many of the strategies designed to minimize the risk of fetal compromise (such as frequent antepartum testing) may have different practical effects in populations with different levels of access to transportation, child care, and appropriate monitoring facilities, this lack of information is disappointing.

Review of national data from birth certificates and hospital discharges suggests that there may be differences in the clinical characteristics of prolonged pregnancy among women in different ethnic and socioeconomic groups. In spite of the multiple limitations of the data, it is striking that two different data sources both show that black women with prolonged pregnancy are more likely to have low birthweight infants than white or Hispanic women. Black women are consistently more likely to have low birthweight infants at other gestational ages as well. Black women also are more likely to have diagnoses of intrauterine growth restriction and oligohydramnios. Women with Medicaid or no insurance are also more likely to have growth restriction and oligohydramnios. We did not explore the degree to which the effects of race might be confounded by economic status, or vice versa, primarily because of problems caused by missing data. Other potential confounders include differences in the use of ultrasound for dating and differences in the use of antepartum testing for prolonged pregnancy. These findings should be investigated further using higher quality data and appropriate epidemiological and statistical methodologies.